University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2004 *Paper* 156

Estimating a Survival Distribution with Current Status Data and High-Dimensional Covariates

Mark J. van der Laan* Aad van der Vaart†

*Division of Biostatistics, School of Public Health, University of California, Berkeley, laan@berkeley.edu

†Free University, Amsterdam, The Netherlands

Estimating a Survival Distribution with Current Status Data and High-Dimensional Covariates

Mark J. van der Laan and Aad van der Vaart

Abstract

We consider the inverse problem of estimating a survival distribution when the survival times are only observed to be in one of the intervals of a random bisection of the time axis. We are particularly interested in the case that high-dimensional and/or time-dependent covariates are available, and/or the survival events and censoring times are only conditionally independent given the covariate process. The method of estimation consists of regularizing the survival distribution by taking the primitive function or smoothing, estimating the regularized parameter by using estimating equations, and finally recovering an estimator for the parameter of interest.

1. Introduction

Under current status type censoring the time of occurrrence T of an event of interest is never observed, but instead at a random monitoring time C it is observed if the event has occurred or not. In the first case the indicator $\Delta = 1\{T \leq C\}$ takes the value 1; it is 0 otherwise. Throughout the paper we assume that we observe a random sample from the distribution of (C, Δ, L), where L is a covariate process. We focus on the estimation of the marginal survival function $S(t) = \mathrm{P}(T > t)$ at a fixed point $t > 0$.

In the simplest version of the current status model the time of interest T and the monitoring time C are assumed independent and S can be estimated by the nonparametric maximum likelihood estimator $\hat{S}$, defined as the maximizer of the likelihood

$$S \mapsto \prod_{i=1}^{n} (1 - S(C_i))^{\Delta_i} S(C_i)^{1-\Delta_i} \tag{1.1}$$

over all survival distributions S. The asymptotic distribution of $\hat{S}(t)$ was first obtained by Groeneboom (1987), who shows that $n^{1/3}\big(\hat{S}(t) - S(t)\big)$ converges to a nontrivial limiting law, which can be derived as a functional of Brownian motion. We refer to Groeneboom and Wellner (1992) and Van der Vaart and Wellner (1996, Example 3.2.15) for published derivations of this result (using somewhat different methods of proof). An attractive feature of the maximum likelihood estimator is that it does not require the specification of a bandwidth parameter. A drawback is its slow rate of convergence, but this cannot be improved without a-priori restrictions on the parameters. In this paper we aim at constructing estimators in more complicated current status models that retain the good properties of the maximum likelihood estimators.

Often the assumption of independence of T and C is unrealistic, but can be replaced by the assumption of conditional independence given a covariate process L. (In the case of a time-varying covariate process this is to be interpreted as "independent at every time instant given the past". See Section 3.) Furthermore, an observed covariate process may be used to improve the efficiency of estimation of $S(t)$. One possibility is to model the survival time T given the covariate L by a Cox model and use the method of maximum likelihood. For a time-independent covariate vector L, this entails maximizing the likelihood

$$(\theta, \Lambda) \mapsto \prod_{i=1}^{n} \Big(1 - e^{-e^{\theta^T L_i} \Lambda(C_i)}\Big)^{\Delta_i} \Big(e^{-e^{\theta^T L_i} \Lambda(C_i)}\Big)^{1-\Delta_i},$$

over all vectors θ and cumulative hazard functions Λ. In the continuous case the survival distribution of interest would be found as $S(t) = \mathrm{E}_L e^{-e^{\theta^T L} \Lambda(t)}$ and could be estimated by plugging in the estimators for θ and Λ. This model is studied by Huang (1996) and Murphy and Van der Vaart (1997), who prove that, under conditions and if the Cox model is correctly specified, then the maximum likelihood

estimator for (θ, Λ) is consistent, $\hat{\theta}$ converges at $\sqrt{n}$-rate with a normal limit distribution and $\hat{\Lambda}$ has a $n^{1/3}$-rate in L_2-norm. The limiting behaviour of $n^{1/3}(\hat{\Lambda} - \Lambda)(t)$ and $n^{1/3}(\hat{S} - S)(t)$ at a fixed point t appears to be unknown.

In the set-up of Huang (1996) the distribution of C is modelled nonparametrically (i.e. is left completely unspecified), whereas the distribution of T is modelled semiparametrically, by the Cox model. If the Cox model fails, then the resulting estimators will be inconsistent. One purpose of the present paper is to reverse the modelling assumptions: our estimators work for any choice of the distribution of T, if a correct model for the distribution of the observation times C is available, for instance a Cox model. Here we follow the method introduced by Robins (1993), Robins and Rotnitzky (1992), and Robins and Van der Laan (1998). From a practical perspective it may be more reasonable to put the modelling assumptions on the distribution of the observation times. As these are observed the experimenter may be more able to formulate a model and check its goodness-of-fit. In certain situations (e.g. animal sacrifice studies) the observation times may be under the control of the experimenter and hence will even have a known distribution.

We are especially interested in the situation that the covariate vector or process L is high-dimensional or time-dependent. Then some modelling appears to be necessary to make the estimation problem feasible, due to the "curse of dimensionality". For instance, a full nonparametric likelihood in the case of a time-independent covariate vector L would take the form

$$\prod_{i=1}^{n} \big(F(C_i|\, L_i)\big)^{\Delta_i} \bar{F}(C_i|\, L_i)^{1-\Delta_i} G\big(\{C_i\}|\, L_i\big).$$

Here F and G denote the conditional distribution functions of T and C given L. For maximum likelihood estimation of $S(t) = 1 - \mathrm{E}F(t|\, L)$ we could drop the term involving G, but it still appears to be unfeasible to maximize the resulting likelihood without making some severe restrictions on the shape of $F(t|\, l)$, in particular regarding its dependence on the (high-dimensional) argument l. For a discussion of this issue see for instance the discussions of the paper Murphy and Van der Vaart (2000), and Robins and Ritov (1997). This situation aggravates if the covariate process is high-dimensional and/or time-dependent.

Robins and Van der Laan (1998) have suggested and implemented estimators of $S(t)$ based on estimating equations, thus avoiding the curse of dimensionality of the likelihood. The basic idea goes back to Robins (1993) and Robins and Rotnitzky (1992). The method requires that the conditional distribution $G(c|\, l)$ of the censoring times be estimated consistently (at some rate) and then works for any $F(t|\, l)$, or alternatively makes this assumption with the roles of F and G reversed. The curse of dimensionality is avoided by focusing on the estimation of a low-dimensional parameter, such as the marginal distribution $S(t)$, letting the estimation of the other parameters intervene at most in preliminary steps, and not striving after full, theoretical efficiency.

As is clear from the results mentioned previously, the estimation of $S(t)$ based on current status data is an inverse problem, with an optimal rate of convergence of $n^{1/3}$ if F is specified nonparametrically or through a Cox model. Because standard theory for estimators defined by estimating equations always yields $\sqrt{n}$-rates of convergence, our use of estimating equations needs explanation. Our approach is to use estimating equations for a regularized version of the parameter $S(t)$. Here we consider three types of regularization. The first is to smooth $S(t)$ by a kernel with bandwidth converging to zero; this was already suggested in Robins and Van der Laan (1998). Whereas Robins and Van der Laan (1998) state results for fixed bandwidths, in the present paper we validate the method for bandwidths converging to zero. The second method of regularization is to estimate the primitive function $\mathbb{S}$ of S, replace this estimator by its least concave majorant (as S is decreasing, its primitive function is concave), and finally estimate S by the derivative of the majorant. This method could also be described as isotonization of a naive, preliminary estimator of S. The third method combines the two methods and consists of isotonization of the estimator of the smoothed S. The second method is perhaps preferable, because it is bandwidth-independent and produces monotone estimators of S.

In order to identify S from the data it is necessary to make assumptions on the dependence structure of T, C and the covariate process L. In the case that L is time-independent we assume that T and C are conditionally independent given L. More generally, if L is a stochastic process, we assume the coarsening-at-random assumption. Informally, this requires that at every time point t the intensity of the occurrence of C depends only on the history of T and the covariate process up to that time point. The assumption is described in more detail in Section 3.

The organization of the paper is as follows. In Section 2 we single out the case of time-independent covariates. In Section 3 the method is shown to work in the greater generality (and complexity) of "coarsening at random" with time-dependent covariates. Sections 4 and 5 contain results about concave majorants and entropies that are used in the proofs of the results.

Throughout we assume that the kernel k used for smoothing is a bounded probability density on $[-1, 1] \subset \mathbb{R}$ with mean zero. The corresponding survival function is denoted $\bar{K}$. We let P, or $P_{F,g}$ to stress dependence on parameters, denote the true distribution of a single observation $(C, 1_{T \leq C}, L)$ or $(C, 1_{T \leq C}, L^C)$ in the time-dependent case, $\mathbb{P}_n$ the empirical measure of the observations and $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ the empirical process. The notation $a \lesssim b$ means $a \leq Db$ for a positive constant D that is fixed throughout. Definitions and notations of entropy and covering numbers are given in Section 5.

2. Time-independent Covariates

The case that the covariate process is a random vector L that is conceptually simpler. For this reason we single out this special case in this section. Throughout this section the variables T and C are assumed to be conditionally independent given a p-dimensional vector L, with conditional distribution functions $F(t|\, l)$ and $G(c|\, l)$, respectively, where $G(\cdot|\, l)$ possesses a positive Lebesgue density $g(\cdot|\, l)$ (at least in a neighbourhood of t). Let $\bar{F}(t|\, l)$ and $\bar{G}(c|\, l)$ denote the corresponding conditional survival functions, and let S be the marginal survival function corresponding to $F(t|\, l)$, i.e. $S(t) = \mathrm{E}_L \bar{F}(t|\, L)$, where E_L means "taking the expectation relative to L".

Our estimating equations are based on the function

$$\psi(F,g,r)(c,\delta,l) = \frac{r(c)\big(F(c|\, l) - \delta\big)}{g(c|\, l)} + \int_0^\infty r(s)\bar{F}(s|\, l)\, ds. \tag{2.1}$$

Up to a constant this is the efficient influence function for estimating the functional $\int_0^\infty r(s)S(s)\, ds$ for a given function $r\colon [0,\infty) \to \mathbb{R}$, in the model in which the conditional distribution $F(t|\, l)$ and the marginal distribution of L are completely unspecified. In the special case that r is constant and the covariate vector is empty this is proved in Van der Vaart (1991a), who also gives precise conditions for the parameter $\int_0^\infty r(s)S(s)\, ds$ to be a differentiable functional on the model. The present more general formula follows by similar arguments and is given in Robins and Van der Laan (1998, formula (10)). It can also be deduced from the more general formula for time-dependent covariates given in Section 3. Clearly some conditions are necessary to make the formula well defined and represent the efficient influence function. Sharp conditions need not concern us here. As we shall be interested in applying the formula with r a local kernel, it will suffice for us that the function $g(\cdot|\, l)$ be bounded away from zero in a neighbourhood of t. The fact that (2.1) gives the efficient influence function is important for the interpretation of our estimators, but does not intervene in the proofs of our results.

By direct computation it follows that, for any F_1, F, g_1, g,

(2.2)

$$P_{F,g}\psi(F,g,r) = \int_0^\infty r(s)S(s)\, ds$$

$$P_{F,g}\big(\psi(F_1,g_1,r) - \psi(F,g,r)\big) = \mathrm{E}_L \int r(c)\big(F_1(c|\, L) - F(c|\, L)\big)\Big(\frac{g(c|\, L)}{g_1(c|\, L)} - 1\Big)\, dc.$$

By the first equation we can view $\mathbb{P}_n\psi(F_n,g_n,r)$ as an estimator of $\int_0^\infty r(s)S(s)\, ds$. The second equation shows that an estimating equation based on ψ is "unbiased" as soon as either F or g is correctly specified.

2.1. Smoothing

Our first estimator of $S(t)$, for a fixed value of t, is

$$S_{n,b}(t) = \mathbb{P}_n\psi(F_n,g_n,k_{b,t}),$$

where F_n and g_n are preliminary estimators of F and g, and $k_{b,t}$ is a kernel of bandwidth $b = b_n$ centered at t, defined as $k_{b,t}(s) = k((s-t)/b)/b$ for k a probability density supported on $[-1, 1]$. The estimator $S_{n,b}(t)$ can be viewed as an estimator of $P_{F,g}\psi(F, g, k_{b,t}) = \int_0^\infty k_{b,t}(s)S(s)\,ds$, which approaches $S(t)$ as $b \to 0$.

The preliminary estimators F_n and g_n need to be chosen such that, for some F_∞ and g_∞,

$$\int_{t-b_n}^{t+b_n} \mathrm{E}_L\big(g_n(c|\,L) - g_\infty(c|\,L)\big)^2\,dc = o_P(b_n). \tag{2.3}$$

$$\int_{t-b_n}^{t+b_n} \mathrm{E}_L\big(F_n(c|\,L) - F_\infty(c|\,L)\big)^2\,dc = o_P(b_n). \tag{2.4}$$

$$\int_{t-b_n}^{t+b_n} \big|\mathrm{E}_L\big(F_n(c|\,L) - F(c|\,L)\big)\big(g_n(c|\,L) - g(c|\,L)\big)\big|\,dc = o_P(b_n^2). \tag{2.5}$$

The expectation E_L in these conditions refers to the integrable over the variable L visible in the formulas. The index P in the order terms on the right sides refers to the n observations on which the estimates F_n and g_n are based, which are not visible in the notation.

The limits F_∞ and g_∞ in (2.3)-(2.4) may be arbitrary, but the third requirement (2.5) suggests that at least one of F_∞ or g_∞ must be equal to the "true" value, F or g, of the corresponding parameter. Note that the integration intervals are of length $2b_n$, so that the first two displays do not really require a rate of convergence of the estimators, but would typically be satisfied if g_n and F_n are ("locally uniformly") consistent for g_∞ and F_∞. By the same argument equation (2.5) requires roughly that its integrand be $o_P(b_n)$.

In our typical application we achieve these conditions by constructing g_n to converge to the true value g at rate $o_P(b_n)$ for $b_n = b_1 n^{-1/3}$, and construct F_n to be merely consistent for some F_∞.

Apart from these consistency requirements we also restrict the sizes of the ranges of the estimators, as measured by entropy, so that we can use empirical process methods to handle the estimated parameters F_n and g_n in the definition of $S_{n,b}$. Let $N = I \times \mathbb{R}^p$ for some fixed neighbourhood I of t. Then we assume that there exist $\eta > 0$ and classes $\mathcal{F}_n$ and $\mathcal{G}_n$ of functions $f\colon [0,\infty) \times \mathbb{R}^p \to [0, 1]$ and $g\colon [0,\infty) \times \mathbb{R}^p \to [\eta, 1/\eta]$ such that with probability tending to one $F_n 1_N$ is contained in $\mathcal{F}_n$ and $g_n 1_N$ is contained in $\mathcal{G}_n$, as $n \to \infty$, and such that, for some $V < 2$,

$$\sup_Q \log N\big(\varepsilon, \mathcal{F}_n 1_N, L_2(Q)\big) \lesssim \left(\frac{1}{\varepsilon}\right)^V, \tag{2.6}$$

$$\sup_Q \log N\big(\varepsilon, \mathcal{G}_n 1_N, L_2(Q)\big) \lesssim \left(\frac{1}{\varepsilon}\right)^V. \tag{2.7}$$

THEOREM 2.1. *Assume that (2.3)-(2.5) and (2.6)-(2.7) are satisfied, where $F_\infty(\cdot|\,l)$, $F(\cdot|\,l)$, $g_\infty(\cdot|\,l)$ and $g(\cdot|\,l)$ are continuous at t, uniformly in l, and $g_\infty(\cdot|\,l)$ and $g(\cdot|\,l)$ are*

bounded away from zero and infinity in a neighbourhood of t, uniformly in l. Assume that S is differentiable at t and let $b_n = b_1 n^{-1/3}$. Then $n^{1/3}\big(S_{n,b_n}(t) - S(t)\big)$ converges in distribution to a mean-zero normal distribution with variance $b_1^{-1}\sigma^2 \int k^2(s)\,ds$, where

$$\sigma^2 = \mathrm{E}_L\big[F(t|\,L)\bar{F}(t|\,L) + \big(F(t|\,L) - F_\infty(t|\,L)\big)^2\big]\frac{g(t|\,L)}{g_\infty(t|\,L)^2}. \tag{2.8}$$

Proof. We can decompose the difference $S_{n,b}(t) - S(t)$ as a sum of three terms

$$\begin{aligned}(\mathbb{P}_n - P_{F,g})\psi(F_n, g_n, k_{b,t}) + P_{F,g}\big(\psi(F_n, g_n, k_{b,t}) - \psi(F, g, k_{b,t})\big)\\ + \big(P_{F,g}\psi(F, g, k_{b,t}) - S(t)\big).\end{aligned}$$

We shall show that the second and third terms are $o_P(b)$, whereas $\sqrt{nb}$ times the first term is asymptotically normal with mean zero and variance σ^2 times the squared L_2-norm of the kernel.

The third term can be rewritten as

$$\mathrm{E}_L \int_0^\infty k_{b,t}(s)\bar{F}(s|\,L)\,ds - S(t) = \int_0^\infty k_{b,t}(s)S(s)\,ds - S(t).$$

This can be seen to be $o(b)$ by Taylor expansion of S around t, for every mean-zero kernel k with $\int |s|k(s)\,ds < \infty$ and every S that is differentiable at t. The argument is the usual one for controlling the bias of a kernel estimator.

The absolute value of the second term can be rewritten as

$$\begin{aligned}&\Big|\int k_{b,t}(s)\mathrm{E}_L\big(F_n(c|\,L) - F(c|\,L)\big)\Big(\frac{g(c|\,L)}{g_n(c|\,L)} - 1\Big)\,dc\Big|\\ &\qquad \lesssim \frac{1}{b}\int_{t-b_n}^{t+b_n}\Big|\mathrm{E}_L\big(F_n(c|\,L) - F(c|\,L)\big)\Big(\frac{g(c|\,L)}{g_n(c|\,L)} - 1\Big)\Big|\,dc,\end{aligned}$$

in view of the compact support and boundedness of k. The random functions g_n are with probability tending to one bounded below by $\eta > 0$ on a neighbourbood of t. Therefore, the right side of the preceding display is $o_P(b)$ by assumption (2.5).

We conclude the proof by proving that the sequence $\sqrt{b}\,\mathbb{G}_n\psi(F_n, g_n, k_{b,t})$ converges in distribution to a normal distribution with mean zero and variance $\sigma^2\|k\|_2^2$, as claimed. The influence function ψ in (2.1) is a sum of two terms, which we denote by

$$\begin{aligned}\psi_1(F, g, r)(c, \delta, l) &= \frac{r(c)\big(F(c|\,l) - \delta\big)}{g(c|\,l)}\\ \psi_2(F, g, r)(c, \delta, l) &= \int_0^\infty r(s)\bar{F}(s|\,l)\,ds.\end{aligned} \tag{2.9}$$

We first show that the second term does not give a contribution to the limit distribution.

The functions $F_n(c|\,l)1_N(c, l)$ are by assumption with probability tending to one contained in a deterministic class $\mathcal{F}_n$ whose uniform entropy is of the order $(1/\varepsilon)^V$, for

some $V < 2$, relative to the envelope function 1, uniformly in n. Then the functions $k_{b,t}(c)F_n(c|\,l)$ are contained in the class $k_{b,t}\mathcal{F}_n$ and this has uniform entropy of the same order relative to the envelope function $k_{b,t}$, in view of Lemma 5.1 (below). Next Lemma 5.2, applied with $t = r = 2 \geq s \geq 1$ and R the uniform measure on a fixed neighbourhood of t shows that the functions $l \mapsto \int k_{b,t}(s)F_n(s|\,l)\,ds$ are contained in a class of functions $\bar{\mathcal{F}}_n$ that has uniform entropy of the order $(1/\varepsilon)^{2V/s}$ relative to the envelope function $(\bar{k}_{b,t})_s = b^{1/s-1}\|k\|_s$. Choose $s \geq 1$ such that $2V/s < 2$. Then in view of inequality (5.1) we obtain that, with high probability,

$$\left|\mathbb{G}_n \int k_{b,t}(s)F_n(s|\,L)\,ds\right| \leq \sup_{f\in\bar{\mathcal{F}}_n} |\mathbb{G}_n f| = O_P(b^{1/s-1}).$$

Then $\sqrt{b}\,\mathbb{G}_n \int k_{b,t}(s)F_n(s|\,L)\,ds$ converges to zero in probability, if we choose s also to satisfy $1/s - 1 + \frac{1}{2} > 0$. Any s such that $V < s < 2$ satisfies both requirements. Because $\mathbb{G}_n 1 = 0$ we can replace F_n by $\bar{F}_n$ without loss of generality.

Finally, we show that the sequence $\sqrt{b}\,\mathbb{G}_n\psi_1(F_n, g_n, k_{b,t})$ converges in distribution to a normal distribution with mean zero and variance $\sigma^2\|k\|_2^2$. We can obtain the functions $(\delta - F(c|\,l))/g(c|\,l)1_N(c,l)$ as a Lipschitz transformation (in the sense of Lemma 5.1 below) applied to the class of all $F \in \mathcal{F}_n$, the class of all $g \in \mathcal{G}_n$ and the class consisting of the single function δ, where we allow only functions g that are bounded away from zero and infinity. Next we obtain the functions $\psi_1(F, g, k_{b,t})$ by multiplication by the single function $k_{b,t}$. In view of the assumptions and Lemma 5.1 there exist deterministic classes $\mathcal{H}_n$ of functions that contain the random functions $\psi_1(F_n, g_n, k_{b,t})$ with probability tending to one and that have uniform entropy of the order $(1/\varepsilon)^V$ relative to the envelope function a multiple of $k_{b,t}$, uniformly in n. Here

$$P_{F,g}(\sqrt{b}k_{b,t})^2 = \int k^2(x)\mathrm{E}_L g(t + bx|\,L)\,dx,$$

and

$$P_{F,g}(\sqrt{b}k_{b,t})^2 1_{\sqrt{b}k_{b,t}\geq\varepsilon\sqrt{n}} = 0,$$

as soon as $\|k\|_\infty < \varepsilon\sqrt{nb}$. Thus the envelope functions satisfy the Lindeberg condition. It now follows that $\sqrt{b}\,\mathbb{G}_n\psi_1(F', g', k_{b,t})$ is asymptotically tight as a process indexed by (F', g') varying over the class as just described. (Cf. Van der Vaart and Wellner (1996), Theorem 2.11.22.) The desired result, that $\sqrt{b}\,\mathbb{G}_n\psi(F_n, g_n, k_{b,t})$ is asymptotically normal as claimed, follows provided it can be shown that

$$bP_{F,g}\big(\psi_1(F_n, g_n, k_{b,t}) - \psi_1(F_\infty, g_\infty, k_{b,t})\big)^2 \xrightarrow{\mathrm{P}} 0,$$

and

$$b\,\mathrm{var}\,\psi_1(F_\infty, g_\infty, k_{b,t}) \to \sigma^2\|k\|_2^2.$$

(See Van der Vaart and Wellner (1996), Chapter 3.11.) Because the random functions g_n are bounded away from zero with probability tending to one and the functions F_n

are bounded by 1, the second moment in the first display can be bounded up to a constant by

$$b\mathrm{E}_L \int k_{b,t}^2(s)\Big[\big(F_n(s|\,L) - F_\infty(s|\,L)\big)^2 + \big(g_n(s|\,L) - g_\infty(s|\,L)\big)^2\Big]\,dc + o_P(1).$$

This converges to zero in probability by assumption, since $k_{b,t} \lesssim (1/b)1_{[t-b_n,t+b_n]}$. Next we have that

$$\begin{aligned} b\mathrm{E}\psi_1(F_\infty, g_\infty, k_{b,t})^2 &= b\mathrm{E}_{L,C}\mathrm{E}\big[\big(\Delta - F_\infty(C|\,L)\big)^2|\,C, L\big]\,\frac{k_{b,t}^2(C)}{g_\infty^2(C|\,L)} \\ &= \int bk_{b,t}^2(c)\mathrm{E}_L\big[F(c|\,L)\bar{F}(c|\,L) + \big(F(c|\,L) - F_\infty(c|\,L)\big)^2\big]\,\frac{g(c|\,L)}{g_\infty^2(c|\,L)}\,dc. \end{aligned}$$

This converges to σ^2. Combined with the fact that $b(\mathrm{E}\psi_1(F_\infty, g_\infty, k_{b,t}))^2 \to 0$, this gives the desired result. ∎

2.2. Isotonization

Our second estimator is based on isotonization. As explained in Robertson, Wright and Dykstra (1988) it is fruitful to visualize the isotonization through the process of computing and differentiating the least concave majorant of a primitive function. Let $\bar{K}_{0,t} = 1_{[0,t)}$ be the survival function of the Dirac measure at t, and consider the process

$$\mathbb{S}_{n,0}(t) = \mathbb{P}_n\psi(F_n, g_n, \bar{K}_{0,t}),$$

given initial estimators F_n and g_n of F and g. For each fixed t this can be considered an estimator of

$$\mathbb{S}(t) := P_{F,g}\psi(F, g, \bar{K}_{0,t}) = \int_0^t S(s)\,ds,$$

a primitive function of S. Because S is decreasing, its primitive function is concave. Therefore, we may expect to improve the estimator $\mathbb{S}_{n,0}(t)$ by replacing this function by its least concave majorant. The left derivative $\hat{S}_{n,0}$ of this least concave majorant is decreasing and, if evaluated at t, is our second estimator for $S(t)$. (The choice for left- rather than right derivative is for definiteness and has no particular advantage.) As $t \mapsto \mathbb{S}_{n,0}(t)$ is the sum of a step function and a concave function, the concave majorant and hence our estimator are easy to compute by any of the standard algorithms for isotonic regression.

In many situations the functions $g(c|\,l)$ or $F(c|\,l)$ may not be estimable on the whole interval $[0,\infty)$. Therefore, it is preferable to define the least concave majorant relative to a general subinterval $I \subset [0,\infty)$, that contains t in its interior. Thus we define the estimator $\hat{S}_{n,0}(t)$, more generally, as the left derivative at t of the least concave majorant of the function $\mathbb{S}_{n,0}\colon I \to \mathbb{R}$ on a fixed, compact neighbourhood I of t.

The preliminary estimators F_n and g_n need to be chosen such that, for all $M > 0$, all sufficiently small $\delta > 0$ and some F_∞ and g_∞,

$$\int_{t-Mb_n}^{t+Mb_n} \mathrm{E}_L\big(g_n(c|\,L) - g_\infty(c|\,L)\big)^2\,dc = o_P(b_n). \tag{2.10}$$

$$\int_{t-Mb_n}^{t+Mb_n} \mathrm{E}_L\big(F_n(c|\,L) - F_\infty(c|\,L)\big)^2\,dc = o_P(b_n). \tag{2.11}$$

$$\int_{t-Mb_n}^{t+Mb_n} \big|\mathrm{E}_L\big(F_n(c|\,L) - F(c|\,L)\big)\big(g_n(c|\,L) - g(c|\,L)\big)\big|\,dc = o_P(b_n^2). \tag{2.12}$$

$$\mathrm{E}\int_{t-\delta}^{t+\delta} \big|\mathrm{E}_L\big(F_n(c|\,L) - F(c|\,L)\big)\big(g_n(c|\,L) - g(c|\,L)\big)\big|\,dc \lesssim b_n(\delta \vee b_n), \tag{2.13}$$

$$\int_I \big|\mathrm{E}_L\big(F_n(c|\,L) - F(c|\,L)\big)\big(g_n(c|\,L) - g(c|\,L)\big)\big|\,dc = o_P(1). \tag{2.14}$$

In (2.13) the expectation E outside the integral refers to the observations on which the estimators F_n and g_n are based, which are suppressed from the notation. The condition may be relaxed to the assumption that the expectations restricted to sets A_n with $\mathrm{P}(A_n) \to 1$ satisfy the inequality.

THEOREM 2.2. *Assume that (2.10)-(2.14) and (2.6)-(2.7) are satisfied, where the functions $F_\infty(\cdot|\,l)$, $F(\cdot|\,l)$, $g_\infty(\cdot|\,l)$ and $g(\cdot|\,l)$ are continuous at t, uniformly in l, and $g_\infty(\cdot|\,l)$ and $g(\cdot|\,l)$ are bounded away from zero and infinity on I, uniformly in l. Assume that S is differentiable at t with $S'(t) < 0$. Then the sequence $n^{1/3}\big(\hat{S}_{n,0}(t) - S(t)\big)$ converges in distribution to*

$$-S'(t)\operatorname*{argmax}_{u\in\mathbb{R}}\big\{\sigma Z_0(u) + \tfrac{1}{2}S'(t)u^2\big\},$$

where Z_0 is a standard Brownian motion process and σ^2 is given in (2.8).

We defer the proof of this theorem to the next subsection, as it is almost identical to the proof for our third estimator. There we also compares the limit distributions obtained in Theorems 2.1 and 2.2.

The conditions for the present theorem are a little stronger than those of Theorem 2.1. They could be relaxed if we let the interval I relative to which we define the isotonization shrink to t at an appropriate bandwidth. An advantage of using a fixed interval I is that a bandwidth need not be specified a-priori.

One consequence of using a fixed interval is that the isotonization procedure also works under weaker smoothness conditions on S. In particular, if rather than existence of $S'(t)$ we require that, for some fixed $\alpha \in (1, 2)$,

$$\mathbb{S}(t+u) - \mathbb{S}(u) - S(t)u = u^\alpha S^{(\alpha)}(t) + o(|u|^\alpha),$$

as $u \to 0$, for some negative number $S^{(\alpha)}(t)$, then the sequence $n^{(\alpha-1)/(2\alpha-1)}(\hat{S}_{n,0} - S)(t)$ will converge to a nontrivial limit distribution. We could obtain similar rates of

convergence for the smoothed estimator $S_{n,b}$, but only if the bandwidth b is adapted to the new smoothness condition (e.g. the value of α). Another consequence is a faster rate of convergence than $n^{1/3}$ if $S'(t)$ exists, but vanishes. Again this could be achieved by a smoothing method only if the bandwidth is appropriately adjusted.

2.3. Isotonization and Smoothing

Our third estimator combines smoothing and isotonization. Letting $\bar{K}_{b,t}$ be the function $s \mapsto \bar{K}((s-t)/b)$ for a survival function $\bar{K}$ (and $b > 0$), we define

$$\mathbb{S}_{n,b}(t) = \mathbb{P}_n \psi(F_n, g_n, \bar{K}_{b,t}).$$

Next we define $\hat{S}_{n,b}(t)$ to be the left derivative at t of the least concave majorant of $\mathbb{S}_{n,b}$ relative to a fixed, compact neighbourhood I of t, as before. We shall show that this yields a good estimator of $S(t)$ if $b \downarrow 0$ appropriately. Note that we can consider $\bar{K}_{0,t} = 1_{[0,t)}$, employed in the preceding subsection, to be the limit of the kernels $\bar{K}_{b,t}$ as $b \downarrow 0$, as the notation suggests.

THEOREM 2.3. *Assume that (2.10)-(2.14) and (2.6)-(2.7) are satisfied, where the functions $F_\infty(\cdot\,|\,l)$, $F(\cdot\,|\,l)$, $g_\infty(\cdot\,|\,l)$ and $g(\cdot\,|\,l)$ are continuous at t, uniformly in l, and $g_\infty(\cdot\,|\,l)$ and $g(\cdot\,|\,l)$ are bounded away from zero and infinity in I, uniformly in l. Assume that S is differentiable at t with $S'(t) < 0$, and let $b_n = n^{-1/3}$. Then the sequence $n^{1/3}\big(\hat{S}_{n,b}(t) - S(t)\big)$ converges in distribution to*

$$-S'(t) \operatorname*{argmax}_{u \in \mathbb{R}} \big\{ \sigma Z_K(u) + \tfrac{1}{2} S'(t) u^2 \big\},$$

where σ^2 is given by (2.8) and Z_K is a Gaussian process, zero at zero, with continuous sample paths and stationary increments and variance function

$$\mathrm{E} Z_K^2(u) = \int \big(K(s+u) - K(s)\big)^2 \, ds.$$

The process Z_K has in common with Brownian motion that it is continuous, zero at zero, centered and possesses stationary increments. However, unlike the increments of Brownian motion the increments of Z_K are not independent. The process Z_K reduces to Brownian motion only for the degenerate kernel $K = 1_{[0,\infty)}$. This is the limiting case of using a bandwidth $b = 0$ (and an arbitrary kernel), which is excluded in the preceding theorem. Apart from the different process, Z_K or Brownian motion, the limiting distributions in Theorems 2.2 and 2.3 are identical in form. The normal limit distribution of Theorem 2.1 can also be written in this form, with Z the linear process $Z(u) = u\|k\|_2 \xi / \sqrt{b_1}$ for a standard normal variable ξ. It can be shown from this that the limit distribution in Theorem 2.3 is always more concentrated about zero than the limit distribution in Theorem 2.1 for $b_1 = 1$. (See Van der Laan and Van der Vaart (2000).)

This would indicate that application of both smoothing and isotonization yields better estimators. However, it appears that the concentration of the limit distributions in the three theorems can only give moderate guidance regarding the relative quality of the estimators. For instance, it can be shown that the concentration of the estimator in Theorem 2.3 increases indefinitely when using kernels K obtained by rescaling a fixed kernel and letting the scale parameter tend to infinity (yielding undersmoothing) (See Van der Laan and Van der Vaart (2000)). The same phenomenon occurs in Theorem 2.1, where the asymptotic variance tends to zero as $b_1 \to \infty$. It is due to the disappearance of the bias of the estimators in the present form of asymptotics. A more refined asymptotic analysis should balance balance bias and variance, but appears difficult to carry out.

Proof of Theorems 2.2 and 2.3. Define δ to be zero for the proof of Theorem 2.2, and define it to be equal to $b_n = n^{-1/3}$ for the proof of Theorem 2.3.

The analysis is based on an inverse process corresponding to $\hat{S}_{n,\delta}$, similar to the analysis of the Grenander estimator in Groeneboom (1985). We recall some properties of this process in general in Section 4. The inverse property actually appears to be dependent on the primitive process (in our case $\mathbb{S}_{n,\delta}$) being upper semi-continuous. As $\mathbb{S}_{n,\delta}$ may not be upper semi-continuous, we replace it by its smallest upper semi-continuous majorant $\bar{\mathbb{S}}_{n,\delta}$. As shown in Section 4, this process has, for our purposes, almost identical properties to $\mathbb{S}_{n,\delta}$. In particular, the least concave majorants of $\mathbb{S}_{n,\delta}$ and $\bar{\mathbb{S}}_{n,\delta}$ agree on the interior of I and hence our estimator is also the left derivative of the least concave majorant of $\bar{\mathbb{S}}_{n,\delta}$.

Define the process, indexed by $a \in \mathbb{R}$,

$$V_{n,\delta}(a) = \operatorname*{argmax}_{s \in I} \big\{ \bar{\mathbb{S}}_{n,\delta}(s) - as \big\},$$

where we take the largest value in the case that multiple Define δ to be zero for the proof of Theorem 2.2, and define it to be equal to $b_n = n^{-1/3}$ for the proof of Theorem 2.3. maximizers exist. Then $\hat{S}_{n,\delta}(t) < a$ if and only if $V_{n,\delta}(a) < t$, for any $a \in \mathbb{R}$ and t in the interior of I, by Lemmas 4.1 and 4.2. Consequently, we have that $n^{1/3}(\hat{S}_{n,\delta} - S)(t) < x$ if and only if $V_{n,\delta}\big(S(t) + xb\big) < t$. By the rescaling $s \to t + bu$ in the definition of $V_{n,\delta}$ we have that

$$\begin{aligned}
\hat{u}_n := \frac{1}{b}\big(V_{n,\delta}\big(S(t) + xb\big) - t\big) &= \operatorname*{argmax}_{u \in (I-t)/b} \big\{ \bar{\mathbb{S}}_{n,\delta}(t + ub) - \big(S(t) + xb\big)(t + ub) \big\} \\
&= \operatorname*{argmax}_{u \in (I-t)/b} \big\{ (\bar{\mathbb{S}}_{n,\delta} - \mathbb{S}_\delta)(t + ub) - (\mathbb{S}_{n,\delta} - \mathbb{S}_\delta)(t) \\
&\qquad + \mathbb{S}_\delta(t + ub) - \mathbb{S}_\delta(t) - S(t)ub - xub^2 \big\},
\end{aligned}$$

for any function $t \mapsto \mathbb{S}_\delta(t)$ not depending on u. We use the function, with $\bar{K}_{0,t} = 1_{[0,t)}$ and $\bar{K}_{b,t} = \bar{K}((\cdot - t)/b))$,

$$\mathbb{S}_\delta(t) = P_{F,g}\psi(F, g, \bar{K}_{\delta,t}) = \int_0^\infty \bar{K}_{\delta,t}(s) S(s)\, ds,$$

by (2.2). The remainder of the proof consists of two steps. First (see (1) below) we show that $\sqrt{n/b} = b^{-2}$ times the process in curly brackets converges in distribution in $\ell^\infty[-M, M]$, for every fixed M, to the process

$$u \mapsto W(u) := \sigma Z(u) + \tfrac{1}{2} S'(t) u^2 - xu,$$

where Z is Brownian motion in the situation of Theorem 2.2 and Z is the process Z_K in the case of Theorem 2.3. Second (see (2) below) we show that the argmax $\hat{u}_n$ is bounded in probability. Then, because $(I - t)/b_n \to \mathbb{R}$, we can conclude from the continuous mapping theorem for the argmax functional (e.g. Van der Vaart (1998, Corollary 5.58)), that

$$\hat{u}_n = \frac{1}{b}\big(V_{n,\delta}\big(S(t) + xb\big) - t\big) \rightsquigarrow \operatorname*{argmax}_{u \in \mathbb{R}} W(u) =: \hat{u}.$$

In view of the stationary increments of Z we have that

$$\hat{u} = \operatorname*{argmax}_{u \in \mathbb{R}} \Big\{ \sigma Z(u) + \tfrac{1}{2} S'(t) \Big(u - \frac{x}{S'(t)} \Big)^2 \Big\}$$

is equal in distribution to the variable

$$\begin{aligned} &\operatorname*{argmax}_{u \in \mathbb{R}} \Big\{ \sigma Z\Big(u - \frac{x}{S'(t)} \Big) - Z\Big(-\frac{x}{S'(t)} \Big) + \tfrac{1}{2} S'(t) \Big(u - \frac{x}{S'(t)} \Big)^2 \Big\} \\ &\qquad = \operatorname*{argmax}_{v \in \mathbb{R}} \Big\{ \sigma Z(v) + \tfrac{1}{2} S'(t) v^2 \Big\} + \frac{x}{S'(t)} =: \hat{v} + \frac{x}{S'(t)}. \end{aligned}$$

Thus $\mathrm{P}(\hat{u} \le 0) = \mathrm{P}(-S'(t)\hat{v} \le x)$ and the same with $<$ replacing $\le$. We conclude by the portmanteau theorem that the probability that $n^{1/3}(\hat{S}_{n,\delta} - S)(t) < x$ is asymptotically sandwiched between $\mathrm{P}(-S'(t)\hat{v} < x)$ and $\mathrm{P}(-S'(t)\hat{v} \le x)$, whence the result.

(1). If S is differentiable at t, then $\mathbb{S}$ is twice differentiable at t with first and second derivatives $S(t)$ and $S'(t)$. It follows that

$$\mathbb{S}_0(t + ub) - \mathbb{S}_0(t) - S(t)ub = \tfrac{1}{2} S'(t) u^2 b^2 + o(b^2),$$

uniformly in u ranging over compacta, as $b \to 0$. This remains true if $\mathbb{S}_0$ on the left is replaced by $\mathbb{S}_b$, as we now show. Set $\bar{K}_{\delta,t,t+u} = \bar{K}_{\delta,t+u} - \bar{K}_{\delta,t}$. Then $\bar{K}_{0,t,t+u} = 1_{[t,t+u)}$ if $u \ge 0$ and $-1_{[t+u,t)}$ if $u \le 0$, and $|K_{b,t,t+u}| \le 1_{[t-b,t+u+b]}$ if $u \ge 0$ and $|K_{b,t,t+u}| \le 1_{[t+u-b,t+b]}$ if $u \le 0$, as K is supported on $[-1, 1]$ and $b > 0$. Because $\int \big(K(x + u) - K(x)\big)\, dx = u$ and $\int \big(K(x + u) - K(x)\big)(x - \mu)\, dx = \frac{1}{2} u^2$ for any cumulative distribution function K with mean μ, we can write

$$\begin{aligned} &\mathbb{S}_b(t + ub) - \mathbb{S}_b(t) - S(t)ub - \tfrac{1}{2} S'(t) u^2 b^2 \\ &\qquad = \int \bar{K}_{b,t,t+ub}(s) \big[S(s) - S(t) - S'(t)(s - t) \big]\, ds \\ &\qquad \le \int |\bar{K}_{b,t,t+ub}|(s) \varepsilon |s - t|\, ds, \end{aligned}$$

for any $\varepsilon > 0$ if b is sufficiently small, uniformly in u ranging over compacta. The right side is bounded by $\varepsilon b^2(1+|u|)^2$ and hence the left side is $o(b^2)$ uniformly in $|u| \le M$.

Next, by (2.2),

$$(\mathbb{S}_{n,\delta} - \mathbb{S}_\delta)(t+ub) - (\mathbb{S}_{n,\delta} - \mathbb{S}_\delta)(t) = (\mathbb{P}_n - P_{F,g})\psi(F_n, g_n, \bar{K}_{\delta,t,t+ub}) + \mathrm{E}_L \int \bar{K}_{\delta,t,t+ub}(c)\big(F_n(c|\,L) - F(c|\,L)\big)\Big(\frac{g(c|\,L)}{g_n(c|\,L)} - 1\Big)\,dc.$$

The last term is of the order $o_P(b^2)$ by assumption (2.12). We shall show that $\sqrt{n/b}$ times the first term on the right converges in distribution in $\ell^\infty[-M, M]$ to the process σZ, for every fixed M. To see this we decompose $\psi = \psi_1 + \psi_2$ as in (2.9), where again ψ_2 gives a negligible contribution. The functions

$$\psi_1(F_n, g_n, \bar{K}_{\delta,t,t+ub}) = \bar{K}_{\delta,t,t+ub}(c)\frac{F_n(c|\,l) - \delta}{g_n(c|\,l)}, \qquad |u| \le M,$$

are with probability tending to one contained in a class of functions $\mathcal{H}_{n,M}$ that can be obtained by a Lipschitz transformation (in the sense of Lemma 5.1) of the classes:

- $\mathcal{F}_n$ with envelope 1 and uniform entropy bounded by a constant times $(1/\varepsilon)^V$,
- $\mathcal{G}_n$ with lower and upper envelope η and $1/\eta$ and uniform entropy bounded by a constant times $(1/\varepsilon)^V$,
- the class $\{\delta\}$ with envelope 1 and entropy 0,
- the class $\mathcal{K}_{n,M} = \{\bar{K}_{\delta,t,t+ub}\colon |u| \le M\}$ with envelope $1_{[t-M_1 b, t+M_1 b]}$ for $M_1 = M+1$. This class can be constructed by applying the monotone function $\bar{K}$ to the polynomials $c \mapsto a*c+b$, with $a, b \in \mathbb{R}$ and next subtracting the function $\bar{K}_{\delta,t}$. Hence $\mathcal{K}_{n,M}$ is VC of index at most 4 and has uniform entropy bounded by a constant times $\log(1/\varepsilon)$ relative to any envelope function, uniformly in n and M. (E.g. Van der Vaart and Wellner (1996, Theorem 2.6.7, and Lemma 2.6.15 and 2.6.18(viii).).)

By Lemma 5.1 the uniform entropy of the class $\mathcal{H}_{n,M}$ relative to the envelope function a constant times $H_{n,M} = 1_{[t-M_1 b, t+M_1 b]}$ is bounded by a constant times $(1/\varepsilon)^V$. This envelope divided by $\sqrt{b}$ satisfies the Lindeberg condition, because

$$\frac{1}{b}P_{F,g}H_{n,M}^2 = \frac{1}{b}\int_{t-M_1 b}^{t+M_1 b} \mathrm{E}_L g(c|\,L)\,dc \lesssim M,$$

$$\frac{1}{b}P_{F,g}H_{n,M}^2 1_{H_{n,M}/\sqrt{b} \ge \varepsilon\sqrt{n}} = 0,$$

as soon as $\varepsilon\sqrt{nb} > 1$. It follows that the processes $u \mapsto \mathbb{G}_n\psi_1(F_n, g_n, \bar{K}_{\delta,t,t+bu})/\sqrt{b}$ converge in distribution in $\ell^\infty[-M, M]$ to σZ if

$$\sup_{|u|\le M} \frac{1}{b}P_{F,g}\Big(\psi_1(F_n, g_n, \bar{K}_{\delta,t,t+ub}) - \psi_1(F_\infty, g_\infty, \bar{K}_{\delta,t,t+ub})\Big)^2 \to 0,$$

and

$$\frac{1}{b}P_{F,g}\psi_1(F_\infty, g_\infty, \bar{K}_{b,t,t+ub})\psi_1(F_\infty, g_\infty, \bar{K}_{b,t,t+vb}) \to \sigma^2 \mathrm{E}Z(u)Z(v).$$

This can be verified as in the proof of Theorem 2.1.

The functions $\bar{K}_{\delta,t,t+bu}F_n(c|\,l)$, where u ranges over $[-M,M]$, are with high probability contained in the class $\mathcal{J}_{n,M}$ of functions obtained by taking products of functions from the classes $\mathcal{K}_{n,M}$ and $\mathcal{F}_n$. This class has uniform entropy relative to envelope function $1_{[t-M_1b,t+M_1b]}$ bounded by a multiple of $(1/\varepsilon)^V$. By Lemma 5.2 the functions $\int \bar{K}_{b,t,t+bu}(c)F_n(c|\,l)\,dc$ are contained in a class of functions $\bar{\mathcal{J}}_{n,M}$ with uniform entropy bounded by a multiple of $(1/\varepsilon)^{2V/s}$ relative to the envelope functions

$$\bar{J}_{n,M}(l) = \Big(\int 1_{[t-M_1b,t+M_1b]}(c)\,dc\Big)^{1/s} = (2M_1b)^{1/s},$$

for any fixed $s \in [1,2]$. For $2V/s < 2$ it follows that

$$\frac{1}{\sqrt{b}} \sup_{|u|\le M} \big|\mathbb{G}_n\psi_2(F_n, g_n, \bar{K}_{\delta,t,t+bu})\big| \le \frac{1}{\sqrt{b}} \sup_{j\in\bar{\mathcal{J}}_{n,M}} |\mathbb{G}_n j| = O_P(b^{1/s-1/2}),$$

which converges to zero in probability if $s < 2$. The two constraints on s are met by any s such that $V < s < 2$.

We now have succeeded to prove the weak convergence of $\sqrt{n/b}$ times the processes $u \mapsto (\mathbb{S}_{n,\delta} - \mathbb{S}_b)(t+ub) - (\mathbb{S}_{n,\delta} - \mathbb{S}_b)(t)$. By Lemma 4.4 this weak convergence is shared by the upper semi-continuous envelopes $u \mapsto (\bar{\mathbb{S}}_{n,\delta} - \mathbb{S}_b)(t+ub) - (\mathbb{S}_{n,\delta} - \mathbb{S}_b)(t)$ of these processes, with the same limit process, as the limit σZ has continuous sample paths. (For the computation of the smallest upper semi-continuous majorant, note that the function $u \mapsto \mathbb{S}_b(t+ub)$ is continuous; furthermore, note that the interval $[t-Mb, t+Mb]$ is contained in I eventually, and the upper semi-continuous majorant of a restriction of a function to a subinterval differs at most at the endpoints of the restriction of the upper semi-continuous majorant on the whole interval.) This concludes the first step of the proof.

(2). The second step is to prove that $\hat{u}_n$ is bounded in probability. Define the processes

$$\mathbb{M}_n(u) = \mathbb{S}_{n,\delta}(t+u) - \mathbb{S}_{n,\delta}(t) - S(t)u - xbu,$$
$$M(u) = \mathbb{S}(t+u) - \mathbb{S}(t) - S(t)u.$$

Then $\hat{u}_n b_n$ maximizes the smallest upper-continuous majorant $\hat{\mathbb{M}}_n$ of $\mathbb{M}_n$ over $u \in I - t$. The moduli of continuity of $\bar{\mathbb{M}}_n$ and $\mathbb{M}_n$ are identical by Lemma 4.5. Because $S'(t) < 0$ and $\mathbb{S}$ is concave, we have that $M(u) \le -c(u^2 \wedge |u|)$ for every $u \in I - t$, and hence $u = 0$ is a unique point of absolute maximum of M. In the situation of Theorem 2.2, for every $\delta > 0$,

$$\mathrm{E} \sup_{\substack{|u|\le\delta\\ u\in I-t}} \big|\sqrt{n}(\mathbb{M}_n - M)(u)\big| \lesssim \mathrm{E} \sup_{\substack{h\in\mathcal{H}_{n,\delta/b}\\ j\in\bar{\mathcal{J}}_{n,\delta/b}}} \big|\mathbb{G}_n(h+j)\big| + \sqrt{n}|x|b\delta$$
$$+ \sqrt{n}\mathrm{E}\int_{c\in I,|c-t|<\delta} \Big|\mathrm{E}_L\big(F_n(c|\,L) - F(c|\,L)\big)\Big(\frac{g(c|\,L)}{g_n(c|\,L)} - 1\Big)\Big|\,dc.$$

With probability tending to one the functions g_n are bounded below by $\eta > 0$. If this is not true with probability one, then we restrict the preceding expectations to

the event where this is true. Then, in view of the maximal inequality (5.1) and the estimates on envelope functions and entropies obtained earlier, this can for sufficiently small $\delta > 0$ be further bounded by

$$\phi_n(\delta) := \sqrt{\delta + b} + (\delta + b)^{1/s} + \sqrt{n}|x|b\delta + \sqrt{n}b(\delta \vee b).$$

For $\delta_n = n^{-1/3}$ we have $\phi_n(\delta_n) \lesssim \sqrt{n}\delta_n^2$. Therefore, it follows by Theorem 3.2.5 of van der Vaart and Wellner (1996) that the maximizer of $u \mapsto \mathbb{M}_n(u)$ possesses a rate of convergence b for the maximizer $u = 0$ of $u \mapsto \mathbb{M}(u)$, provided that it is consistent for zero. To prove the consistency we replace the bound in the preceding display by

$$\sqrt{\delta + b} + (\delta + b)^{1/s} + \sqrt{n}|x|b\delta + \sqrt{n}\eta_n,$$

where by assumption

$$\eta_n := \mathrm{E} \int_I \Big| \mathrm{E}_L\big(F_n(c|\,L) - F(c|\,L)\big)\Big(\frac{g(c|\,L)}{g_n(c|\,L)} - 1\Big)\Big|\, dc \to 0.$$

This bound is valid for every $\delta > 0$ and hence the same Theorem 3.2.5 yields a (suboptimal) rate of convergence of $\hat{u}_n$ to zero (depending on η_n) and hence consistency.

In the situation of Theorem 2.1 the preceding estimates must be augmented by the term

$$\sqrt{n} \sup_{\substack{|u| \le \delta \\ u \in I - t}} \big|\mathbb{S}_b(t + u) - \mathbb{S}_b(t) - \mathbb{S}(t + u) + \mathbb{S}(t)\big|.$$

For sufficiently small δ and $|u| < \delta$

$$\begin{aligned}
&\int (\bar{K}_{b,t,t+u} - \bar{K}_{0,t,t+u})(s) S(s)\, ds \\
&\quad = \int (\bar{K}_{b,t,t+u} - \bar{K}_{0,t,t+u})(s) \big[S(s) - S(t) - (s - t)S'(t)\big]\, ds \\
&\quad \le \int \big(|\bar{K}_{b,t,t+u}| + |\bar{K}_{0,t,t+u}|\big)(s) \varepsilon |s - t|\, ds \le \varepsilon(\delta + b)^2.
\end{aligned}$$

Furthermore, for any $u \in I - t$ the expression is bounded above by

$$\int \big|\bar{K}_{b,t,t+u} - \bar{K}_{0,t,t+u})\big|(s)\, ds,$$

and this converges to zero as $b \to 0$ uniformly in u. ∎

2.4. Preliminary Estimators

All three constructions require initial "estimators" for F and g. These need to satisfy the general conditions given in the preceding sections, but can be chosen in many ways. We indicate some possibilities.

As explained in the introduction our interest is in the situation that L is high-dimensional, in which case it appears to be impossible to obtain good estimators for $F(t|\,l)$ or $g(t|\,l)$ without making some assumptions on the true parameters. In

principle we choose to make the assumption that the true density $g(c|\, l)$ belongs to a given semiparametric model of moderate dimension. One possibility is to assume that the observation times actually are dependent on a low-dimensional subvector of L. In that case we may estimate $g(c|\, l)$ nonparametrically from the data.

Another possibility, which we shall discuss in more detail, is to postulate the Cox model

$$G(t|\, l) = e^{-\Lambda(t)e^{\theta^T l}},$$

where Λ is an unknown cumulative baseline hazard function, and $\theta \in \mathbb{R}^p$. Then a natural estimator $g_n(c|\, l)$ is obtained from the (partial) likelihood estimators $(\hat{\theta}, \hat{\Lambda})$ based on the observations $(C_1, L_1), \ldots, (C_n, L_n)$. We assume that $G(\cdot|\, l)$ possesses a Lebesgue density $g(\cdot|\, l)$, at least in a neighbourhood of t, or, equivalently, that Λ possesses a density λ, and estimate λ by smoothing the maximum likelihood estimator $\hat{\Lambda}_n$. If Λ is twice continuously differentiable, then, with $\delta_n = n^{-1/5}$, we may define

$$g_n(c|\, l) := e^{\hat{\theta}^T l}\hat{\lambda}_n(c)e^{-\hat{\Lambda}_n(t)e^{\hat{\theta}^T l}},$$

where

$$\hat{\lambda}_n(c) = \int k\Big(\frac{s-c}{\delta_n}\Big)\frac{1}{\delta_n}\, d\hat{\Lambda}_n(s).$$

The estimators g_n attains a locally uniform rate of convergence of $n^{-2/5}$ up to a logarithmic factor. (See Andersen, Borgan, Gill and Keiding (1992).) This is certainly enough to satisfy (2.3) and (2.5) or (2.10) and (2.13)-(2.14). We shall also have that $\hat{\lambda}_n'$ converges locally uniformly to λ', and hence $\hat{\lambda}_n$ is with probability tending to one contained in the class of all Lipschitz functions, which has uniform entropy bounded by a multiple of $(1/\varepsilon)$. As $\hat{\Lambda}_n$ ranges over monotone, bounded functions, which also form a class of functions of uniform entropy bounded by a multiple of $(1/\varepsilon)$, condition (2.7) is seen to be satisfied.

Next consider the "estimators" for $F(t|\, l)$. Having ensured the rate conditions (2.5) or (2.13) by construction of $g_n(c|\, l)$ and the modelling assumption on the censoring mechanism, any choice of estimator that is locally uniformly consistent for some F_∞ (not necessarily the true F) and satisfies the entropy condition (2.6) will do.

A computationally attractive procedure, suggested and implemented by Robins and Van der Laan (1998), is to fit the (possibly misspecified) model

$$F_{\alpha,\beta,\gamma}(t|\, l) = \frac{1}{1 + e^{\alpha+\beta t+\gamma^T l}}.$$

This is motivated by the fact that

$$F(t|\, l) = \mathrm{EP}(T \le t|\, C = t, L = l) = \mathrm{E}(\Delta|\, C = t, L = l).$$

Thus we can estimate the parameters (α, β, γ) by fitting a standard logistic regression model to the observed binary outcomes Δ_i given the "covariates" (C_i, L_i). Even if the model is misspecified, the maximum likelihood estimators $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ will typically converges to values $(\alpha_1, \beta_1, \gamma_\infty)$ minimizing a Kullback-Leibler type distance and hence

$F_{\hat{\alpha},\hat{\beta},\hat{\gamma}}$ will converge as well. The entropy condition (2.6) for this type of estimators is easily satisfied, as $F_n(t|\, l)$ ranges over a finite-dimensional set.

More challenging is to construct estimators F_n so as to maximize the asymptotic efficiency of the ensuing estimator for $S(t)$. This efficiency is determined by the parameter σ^2 in (2.8). It is clear from this expression that an optimal choice would be to construct F_n to be consistent for F, in which case (2.8) (with $g_\infty = g$) reduces to

$$\mathrm{E}_L \frac{F(t|\, L)\bar{F}(t|\, L)}{g(t|\, L)}.$$

By the strict concavity of the function $(x, y) \mapsto x(1-x)/y$ on $[0,1] \times (0,\infty)$ and Jensen's inequality this is always bounded above by

$$\frac{F(t)\bar{F}(t)}{g(t)}.$$

The latter expression takes the role of σ^2 in the analogous limiting result for the maximum likelihood estimator of S, defined as in (1.1), in the case that this likelihood is correctly specified, i.e. C and T are unconditionally independent. One conclusion is that in general covariates may help to increase the efficiency of estimating $S(t)$. A second conclusion is that, in case C and T are both unconditionally independent and conditionally independent given L, then our method, which uses the covariate vector L, is equally efficient as the nonparametric maximum likelihood estimator as in (1.1), provided we use an estimator $F_n(t|\, l)$ that is consistent for the true value of F. Thus the method does not loose in efficiency, even though it is consistent under a wide variety of alternative hypotheses.

An attractive way to make these observations operational is to specify a "best guess" of the true F and construct the estimator F_n in such a way that it is consistent if the guess is correct. Here a best guess could consist of the specification of a parametric or low-dimensional semiparametric model, which is thought to contain the true F or to be close to it.

One possibility is to extend the logistic scheme discussed previously to models of the form

$$F(t|\, l) = \frac{1}{1 + e^{\alpha + \beta(t) + \gamma^T l}},$$

where presently β is a function ranging over a class $\mathcal{B}$. Setting $\gamma = 0$, we see that a true F under which T and L are independent will be in the class as soon as $\mathcal{B}$ is large enough. In this case our method will not loose in efficiency relative to the procedure (1.1) as soon as T is independent of L.

Another possibility is to construct F_n from a Cox model

$$\bar{F}(t|\, l) = e^{-e^{\theta^T l}\Lambda(t)}.$$

Presently we do not observe the times T_i, so that the standard Cox estimators cannot be used. However, the parameters (θ, Λ) can be estimated by maximum likelihood

as in Huang (1996). Under the assumption that the Cox model is indeed correct, Huang (1996) and Murphy and Van der Vaart (1997, 1998) prove, under some conditions, that the maximum likelihood estimator is consistent. This can be extended to show that if the true F does not follow the specification of the Cox model, then the maximum likelihood estimator converges to (θ_1, F_∞) determined by the minimum Kullback-Leibler distance of the true model, as can be expected from analogy with likelihood methods for parametric models. The uniform entropy of the class of all $F(t|\, l)$ deriving from Cox models is determined by the uniform entropy of the functions Λ, the part involving l being finite-dimensional. Because the functions Λ are monotone, the entropy condition (2.6) is satisfied with $V = 1$. Again this model contains the submodel in which T and L are independent and T has a completely unspecified distribution.

The θ-component of the maximum likelihood estimator $(\hat{\theta}, \hat{\Lambda})$ is efficient in the semiparametric sense, if the Cox model is correctly specified. (Cf. Huang (1996).) It could be expected that the Λ-component possesses efficiency properties as well. In particular, it could be expected that the maximum likelihood estimator of $S(t)$, which is a smooth transformation of $(\hat{\theta}, \hat{\Lambda})$ has good properties. (We keep these statements vague on purpose, because there is no precise framework for discussing semiparametric "efficiency" in inverse problems. Moreover, the precise behaviour of the maximum likelihood estimators is unknown.) The estimator constructed in the present paper should be expected to be less efficient, even if we follow the procedure using the Cox model outlined in the preceding paragraph. This is the price to be paid for the estimator to have good properties on a bigger model than the Cox model. On the other hand, if we estimate F using a Cox model and this model is correctly specified, then our method would also work even if the model we have used for estimating g is incorrect. This is because the rate conditions (2.5) or (2.13) require only that one of the two preliminary estimators g_n or F_n be consistent. Thus our method may loose in "asymptotic variance", but it will always yield estimators with the correct rate of convergence.

3. Time-dependent Covariates

In this section we allow the covariate L to be time-dependent: $L = (L_t : t > 0)$ is a cadlag stochastic process with values in $\mathbb{R}^p$ (or some more general metric space). We assume that we observe this covariate process up to the random time C. More precisely, we assume that a "survival time" T, a "censoring time" C and a "covariate process" L are defined on a given probability space. Rather than the "full data" (C, T, L) we observe $(C, 1_{T \le C}, L^C)$, where $L^C = (L_{C \wedge t} : t > 0)$ is the process L stopped at time C.

We assume that the censoring mechanism satisfies the *coarsening-at-random* condition. Coarsening at random is a generalization of "missing at random" and was introduced by Heitjan and Rubin (1991) for discrete data, and subsequently developed for general data by Jacobsen and Keiding (1995), and Gill, van der Laan and Robins (1997). In the present setting it entails the assumption that there exists a measurable function, denoted by $g(c \mid t, l)$, of the three arguments $(c, t, l) \in [0, \infty) \times [0, \infty) \times D[0, \infty)^p$ such that

- the function $c \mapsto g(c \mid t, l)$ is a density relative to the Lebesgue measure for every $(t, l) \in [0, \infty) \times D[0, \infty)^p$.
- the function $c \mapsto g(c \mid t, l)$ gives the conditional law of C given $(T, L) = (t, l)$.
- there exists a measurable function $\tilde{g} : [0, \infty) \times D[0, \infty)^p \to [0, \infty)$ such that $g(c \mid t, l) = \tilde{g}(c, l^c)$, where $l^c = (l_{t \wedge c} : t > 0)$.

It can be shown that, given the first two requirements, the third requirement is equivalent to the seemingly weaker assumption that $g(c \mid t, x)$ is a measurable function of the bigger variable $(c, 1_{t \le c}, l^c)$, which is the "observed value", rather than of (c, l^c) only. Thus, informally, CAR requires that the distribution of the censoring mechanism depends on the observed data only. In the case that $L = L_0$ is time-independent, the CAR assumption can be shown to reduce to the assumption of conditional independence of T and C given L_0, made in Section 2.

An intuitive understanding of the coarsening-at-random condition in the present situation with time-dependent covariates can be given in terms of the intensity of the counting process $c \mapsto 1_{C \le c}$. The process

$$M_c = 1_{C \le c} - \int_{[0,c]} 1_{C \ge s} \, d\Lambda(s \mid T, L),$$

where $c \mapsto \Lambda(c \mid T, L)$ is the conditional cumulative hazard function of C given (T, L), is a martingale relative to the filtration $\sigma(1_{C \le s} : s \le c, T, L)$. The Lebesgue density of $\Lambda(\cdot \mid T, L)$ is the conditional hazard or intensity of C and if restricted to the event $C \ge c$ can be interpreted as the the "infinitesimal conditional probability" of the event $C = c$ given $C \ge c$ and given (T, L). Under CAR the conditional density of C given (T, L) can be written in the form $g(c \mid T, L) = \tilde{g}(c, L^c)$ and hence the conditional cumulative hazard function can be written as

$$\Lambda(c \mid T, L) = \int_{[0,c]} \frac{g(s \mid T, L)}{\bar{G}(s- \mid T, L)} \, ds = \int_{[0,c]} \frac{\tilde{g}(s, L^s)}{1 - \int_{[0,s)} \tilde{g}(u, L^u) \, du} \, ds.$$

The last expression shows that the stochastic process $\big(\Lambda(c|\,T,L)\colon c>0\big)$ is adapted to the filtration $\sigma(L^c)=\sigma(L_t\colon t\le c)$. Hence the conditional intensity

$$1_{C\ge c}\frac{\tilde g(c,L^c)}{1-\int_{[0,c)}\tilde g(u,L^u)\,du}$$

of the martingale M_c is $\sigma(C,L^c)$-adapted. Intuitively, the "infinitesimal conditional probability" of the event $C=c$ given $C\ge c$ and given (T,L) depends on c and the oberved covariate process L^c until that time only. In particular, this is the case if the intensity of C at c depends on L_c only.

The preceding observations are important for the motivation and construction of our estimators, but do not intervene in the proof of the main result. The starting point for constructing estimators for $S(t)=\mathrm{P}(T>t)$ is the function

$$\begin{aligned}
&\psi(F,g,r)(c,\delta,l^c)\\
&\quad=\frac{(1-\delta)r(c)}{g(c|\,l)}-\int_0^\infty\Big(\bar F(s|\,l^s)\frac{r(s)}{g(s|\,l)}-\frac{\int_{(s,\infty)}\bar F(u|\,l^u)r(u)\,du}{\bar G(s|\,l)}\Big)\,dM_s^g\\
&\quad=\frac{(1-\delta)r(c)}{g(c|\,l)}-\bar F(c|\,l^c)\frac{r(c)}{g(c|\,l)}+\frac{\int_{(c,\infty)}\bar F(u|\,l^u)r(u)\,du}{\bar G(c|\,l)}\\
&\qquad+\int_{[0,c]}\bar F(s|\,l^s)\frac{r(s)}{\bar G(s|\,l)}\,ds-\int_{[0,c]}\frac{\int_{(s,\infty)}\bar F(u|\,l^u)r(u)\,du}{\bar G^2(s|\,l)}\,dG(s|\,l)\\
&\quad=\frac{(F(c|\,l^c)-\delta)r(c)}{g(c|\,l)}+\int_0^\infty\bar F(s|\,l^s)r(s)\,ds.
\end{aligned}$$

(The last equality follows by partial integration of the fifth term in the middle expression, which yields two terms, the second of which cancels the third and fourth terms.) The process $c\mapsto F(c|\,L^c)$ is a cadlag version of the process $c\mapsto\mathrm{P}(T\le c|\,L^c)$. (It is an assumption that there exists a cadlag version.) By the CAR assumption the function $g(c|\,l)$ depends on its arguments only through (c,l^c). Thus the right side depends only (c,δ,l^c) only, as suggested by the notation on the left side. It is tempting to write $g(c|\,l^c)$ rather than $g(c|\,l)$, or $\tilde g(c,l^c)$ as before, but a conditional (Lebesgue) density at c given a conditioning variable L^c that depends on c is not well-defined, so that we prefer the notation $g(t|\,l)$.

In Robins (1993), van der Laan and Robins (1998), and Van der Vaart (2001) it is shown, under some conditions, that up to a constant this function is the efficient influence function for the parameter $P_{F,g}\mapsto\int_0^\infty S(s)r(s)\,ds$ in the model in which the distribution of the observations is restricted by CAR only, but this again is relevant as background information only.

The last representation of the function ψ is similar to the representation of the influence function in the time-independent case. Estimators of the marginal survival function $S(t)$ can therefore be constructed and analyzed by similar methods. Presently we need preliminary estimators g_n for the conditional density $g(c|\,l)=\tilde g(c,l^c)$ and F_n for the conditional probabilities $F(c|\,L^c)$. As in the time-independent

case it is essential that one of these estimators be consistent for the true parameter, but the other estimator need only stabilize to a limit. This is reassuring, because it will generally be impractical to estimate the conditional probabilities $F(c|\, L^c) = \mathrm{P}(T \leq c|\, L^c)$ consistently, if these really are thought to depend on the covariate process. Because we have full observations on the variables (C, L^C), the estimation of g will be more feasible, even if not easy. We discuss possible preliminary estimators below.

The three types of estimators of the survival probability $S(t)$ take the same forms as in the time-dependent case. Let g_n and F_n be preliminary estimators and let $\mathbb{P}_n$ be the empirical measure of the observations $(C_1, \Delta_1, L_1^{C_1}), \ldots, (C_n, \Delta_n, L_n^{C_n})$. The smoothed estimator of $S(t)$ is

$$S_{n,b}(t) = \mathbb{P}_n \psi(F_n, g_n, k_{b,t}).$$

The isotone estimator is the left derivative of the least concave majorant on an interval I, a neighbourhood of t, of the function

$$t \mapsto \mathbb{S}_{n,0}(t) = \mathbb{P}_n \psi(F_n, g_n, \bar{K}_{0,t}).$$

Finally, the smoothed, isotone estimator is the left derivative of the least concave majorant of the function, with $b = n^{-1/3}$,

$$t \mapsto \mathbb{S}_{n,b}(t) = \mathbb{P}_n \psi(F_n, g_n, \bar{K}_{b,t}).$$

The limiting properties of these estimators can be obtained analogously as in the time-independent case, under analogous conditions. For brevity we give the theorem and the conditions only for the isotone estimator.

The preliminary estimators F_n and g_n should take their values in the collection of all measurable functions $h\colon [0,\infty) \times D[0,\infty)^p \to [0,\infty)$ that depend on their argument (c, l) only through (c, l^c). As in the time-independent case we impose both consistency and entropy conditions.

The estimators need to be chosen such that, for all $M > 0$, all sufficiently small $\delta > 0$ and some F_∞ and g_∞,

$$\int_{t-Mb_n}^{t+Mb_n} \mathrm{E}_L\big(g_n(c|\, L) - g_\infty(c|\, L)\big)^2 \, dc = o_P(b_n). \tag{3.1}$$

$$\int_{t-Mb_n}^{t+Mb_n} \mathrm{E}_L\big(F_n(c|\, L^c) - F_\infty(c|\, L^c)\big)^2 \, dc = o_P(b_n). \tag{3.2}$$

$$\int_{t-Mb_n}^{t+Mb_n} \big|\mathrm{E}_L\big(F_n(c|\, L^c) - F(c|\, L^c)\big)\big(g_n(c|\, L) - g(c|\, L)\big)\big| \, dc = o_P(b_n^2). \tag{3.3}$$

$$\mathrm{E}\int_{t-\delta}^{t+\delta} \big|\mathrm{E}_L\big(F_n(c|\, L^c) - F(c|\, L^c)\big)\big(g_n(c|\, L) - g(c|\, L)\big)\big| \, dc \lesssim b_n(\delta \vee b_n), \tag{3.4}$$

$$\int_I \big|\mathrm{E}_L\big(F_n(c|\, L^c) - F(c|\, L^c)\big)\big(g_n(c|\, L) - g(c|\, L)\big)\big| \, dc = o_P(1). \tag{3.5}$$

Let $N = I \times D[0,\infty)^p$. Then we assume that there exist $\eta > 0$ and classes $\mathcal{F}_n$ and $\mathcal{G}_n$ of functions $f\colon [0,\infty) \times D[0,\infty)^p \to [0,1]$ and $g\colon [0,\infty) \times D[0,\infty)^p \to [\eta, 1/\eta]$ that satisfy the entropy conditions (2.6)-(2.7) such that with probability tending to one $F_n 1_N$ is contained in $\mathcal{F}_n$ and $g_n 1_N$ is contained in $\mathcal{G}_n$, as $n \to \infty$. At first sight the entropy bounds (2.6)-(2.7) may look somewhat complicated, because they use $L_2(Q)$-entropy numbers relative to probability measures Q on the infinite-dimensional space $[0,\infty) \times D[0,\infty)^p$. However, in special cases the estimators g_n or F_n may depend only on lower-dimensional functions of their arguments (c, l^c) and the conditions can be verified by using entropy bounds for sets of functions on finite-dimensional domains.

THEOREM 3.1. *Assume that (3.1)-(3.5) and (2.6)-(2.7) are satisfied, where the maps $c \mapsto F_\infty(c|\, l)$, $c \mapsto F(c|\, l)$, $c \mapsto g_\infty(c|\, l)$ and $g(\cdot|\, l)$ are continuous at t, and $g_\infty(\cdot|\, l)$ and $g(\cdot|\, l)$ are bounded away from zero and infinity on I, uniformly in l. Assume that S is differentiable at t with $S'(t) < 0$. Then the sequence $n^{1/3}\big(\hat{S}_{n,0}(t) - S(t)\big)$ converges in distribution to*

$$-S'(t) \operatorname*{argmax}_{u \in \mathbb{R}} \big\{ \sigma Z_0(u) + \tfrac{1}{2} S'(t) u^2 \big\},$$

where Z_0 is a standard Brownian motion process and

$$\sigma^2 = \mathrm{E}_L \big[F(t|\, L^t) \bar{F}(t|\, L^t) + \big(F(t|\, L^t) - F_\infty(t|\, L^t) \big)^2 \big] \frac{g(t|\, L)}{g_\infty(t|\, L)^2}.$$

Proof. The proof is analogous to the proof of Theorem 2.2. The analogon of equation (2.2) is valid in the form

$$P_{F,g} \psi(F_1, g_1, r) = \int \big(F_1(c) - F(c) \big) r(c) \frac{g(c|\, L)}{g_1(c|\, L)}\, dc + \int_0^\infty \bar{F}_1(s) r(s)\, ds.$$

To see this, we disintegrate the expectation relative to (C, T, L) into the expectations of C given (T, L), of T given L and of L, to see that

$$P_{F,g} \frac{F(C|\, L^C) - 1_{T \le C}}{g(C|\, L)} r(C) = \mathrm{E}_{L,T} \int \big(F_1(c|\, L^c) - F(c|\, L) \big) r(c) \frac{g(c|\, T, L)}{g_1(c|\, L)}\, dc.$$

Next we use the CAR assumption to see that we can drop T from $g(c|\, T, L)$ and that $g(c|\, L)/g_1(c|\, L)$ depends on L^c only. In view of the latter and the orthogonality property of conditional expecations we may next replace $F(c|\, L)$ by $F(c|\, L^c)$, after which the claim follows. ∎

3.1. Preliminary Estimators

In this section we indicate some possibilities for constructing preliminary estimators, where are most interested in preliminary estimators for g. To construct a preliminary estimator for F we may use all available knowledge so as minimize the parameter σ^2 in Theorem 3.1. However, as long as we construct appropriate consistent estimators for g, the estimators F_n need not consistent. One possibility would be to use the

maximum likelihood estimator defined through (1.1), which would mean ignoring all covariate information altogether. Another possibility would be to use only the covariate information available at time zero, through one of the methods discussed in Section 2.4.

The preliminary estimator for g should be a conditional density that satisfies the CAR assumption. The most natural way to model such conditional densities is in terms of the corresponding conditional hazard fuctions $\lambda(c|\, l) = g(c|\, l)/\bar{G}(c|\, l)$. As argued previously the CAR assumption implies that $\lambda(c|\, l)$ depends on its argument (c, l) only through (c, l^c). Conversely, it can be seen by the same argument that if $\lambda(c|\, l)$ depends on (c, l^c) only, then $g(c|\, l) = \lambda(c|\, l)e^{-\Lambda(c|l)}$ satisfies the CAR assumption. Thus any specification

$$\lambda(c|\, l) := \tilde{\lambda}(c, l^c)$$

with $\tilde{\lambda}$ a measurable, nonnegative, locally integrable function leads to a CAR model. The value $\lambda(c, l^c)$ can be interpreted intuitively as the conditional hazard of C at time c given that the covariate process until that time is l^c. If the p covariate processes making up $L \in D[0, \infty)^p$ correspond to measurements of p variables over time, then it can usually be arranged that $\tilde{\lambda}(c|\, l^c)$ depends on l^c only through its final value l_c. Then an $L_2(Q)$-norm on the function $\lambda(c|\, l)$ for a given probability measure Q on $[0, \infty) \times D[0, \infty)^p$ is identical to an $L_2(\tilde{Q})$-norm on the function $\tilde{\lambda}(c, l_c)$ under the measure $\tilde{Q}$ induced on $[0, \infty) \times \mathbb{R}^p$ under the map $(c, l) \mapsto (c, l^c)$. This greatly facilitates the verification of the entropy conditions (2.6)-(2.7).

An attractive approach that illustrates this is to assume the time-dependent Cox model

$$\lambda(c|\, l) = \lambda(c)e^{\theta^T l_c}.$$

We may then estimate the unknown parameters λ (a hazard function) and $\theta \in \mathbb{R}^p$ by the standard (smoothed) Cox likelihood estimators based on the observations $(C_1, L_1^{C_1}), \ldots, (C_n, L_n^{C_n})$. If desired we can also stratify the data and use different base-line hazard functions λ for the different strata, and we may also transform the covariate vector l_c before using it in the linear regression term. In the present case, with $g(c|\, l) = \lambda(c|\, l)e^{-\lambda(c|l)}$, we have

$$\int_{[0,\infty)} \int_{D[0,\infty)^p} g^2(c|\, l)\, dQ(c, l) = \int_{[0,\infty)} \int_{\mathbb{R}^p} \lambda^2(c) e^{2\theta^T l} e^{-\Lambda(c)e^{\theta^T l}}\, d\tilde{Q}(c, l),$$

where $\tilde{Q}$ is the law of (C, L_C) on $\mathbb{R}^{p+1}$ if (C, L) possesses law Q on $[0, \infty) \times D[0, \infty)^p$. Thus the supremum over Q in the entropy conditions (2.6)-(2.7) can be bounded by a supremum over all measures on $\mathbb{R}^{p+1}$ of functions that correspond to a time-independent Cox model. These suprema satisfy the entropy bounds (2.6)-(2.7) under a smoothness assumption on λ, as explained in Section 2.4.

4. Concave Majorants

In this section we list a number of results related to concave and upper semi-continuous majorants that are essential in the proofs of the main results.

Given a function $\Phi\colon I \to \mathbb{R}$ on an interval $I \subset \mathbb{R}$, its least concave majorant $\tilde{\Phi}$ is defined as the pointwise infimum of all concave functions $f\colon I \to \mathbb{R}$ with $f \geq \Phi$. If Φ is upper semi-continuous, then so is the function Φ_a defined by $\Phi_a(s) = \Phi(s) - as$, for any $a \in \mathbb{R}$. If, furthermore, I is compact, then this function attains a maximum on the interval I and the set of points in I at which it is maximal is closed (being the inverse image of the one point set $\{\max \Phi_a\}$ under Φ_a). In that case, we can define

$$\operatorname{argmax} \Phi_a \equiv \operatorname*{argmax}_{s\in I}\big\{\Phi(s) - as\big\} = \max\{s \in I\colon \Phi_a(s) = \max_{t\in I} \Phi_a(t)\}.$$

As a concave, bounded function on I the function $\tilde{\Phi}$ is continuous on the interior of I, is differentiable from the left and the right everywhere in the interior, and is differentiable except at at most countably many points. Let $\tilde{\Phi}'$ be the left derivative of $\tilde{\Phi}$. The following result has been used by several authors, probably first by Groeneboom (1985), and is explicitly stated in Jongbloed (1999). It shows that the functions $\tilde{\Phi}'$ and $a \mapsto \operatorname{argmax} \Phi_a$ are each other's inverses. For a concrete function Φ, such as an empirical distribution function, the validity of the lemma is easily ascertained by a picture. Because we apply the result to functions Φ that in principle may be irregular, we supply a rigorous proof for the general case.

LEMMA 4.1. *Let $\Phi\colon I \subset \mathbb{R}$ be upper semi-continuous on the compact interval $I \subset \mathbb{R}$. Then for any t in the interior of I and any $a \in \mathbb{R}$ we have that $\tilde{\Phi}'(t) < a$ if and only if* $\operatorname{argmax} \Phi_a < t$.

Proof. For simplicity of notation, let the interval I be the unit interval $I = [0, 1]$.

Define a function

$$\ell(t) = \inf_{s<t} \max_{u\geq t} \frac{\Phi(u) - \Phi(s)}{u - s}.$$

Then $\ell(t) < a$ if and only if there exists $s < t$ such that for all $u \geq t$ the quotient in the display is strictly smaller than a, which is equivalent to $\Phi(u) - au < \Phi(s) - as$. (For the "if" we use that the maximum over $u \geq t$ is assumed.) This is equivalent to the largest point of maximum of Φ_a (which exists!) being strictly to the left of t.

If it can be shown that $\ell(t)$ is equal to the left derivative of $\tilde{\Phi}$ at t, then the lemma is proved. We shall show that this is the case, but by somewhat of a detour.

First consider the case that Φ itself is concave. Then the quotients $\big(\Phi(u) - \Phi(s)\big)/(u - s)$ increase to $r(s) := \big(\Phi(t) - \Phi(s)\big)/(t - s)$ as u decreases to t, for fixed s, and $r(s)$ decreases to the left derivative of Φ as $s \uparrow t$. It follows that ℓ is the left-derivative of Φ if Φ itself is concave.

Applying the preceding arguments to $\tilde{\Phi}$ instead of Φ, we see that $\operatorname{argmax}_s\{\tilde{\Phi}(s) - as\} < t$ if and only if $\tilde{\Phi}'(t) < a$. We shall conclude the proof of the lemma by showing that $\operatorname{argmax}_s\{\tilde{\Phi}(s) - as\} < t$ if and only if $\operatorname{argmax}_s\{\Phi(s) - as\} < t$.

Define functions g_t and h_t by

$$g_t(a) = \max_{s\le t, s\in I}\{\Phi(s) - as\}, \qquad h_t(a) = \max_{s\ge t, s\in I}\{\Phi(s) - as\},$$

and define $\tilde g_t$ and $\tilde h_t$ similarly from $\tilde\Phi$ instead of Φ. Then $g_t(a) \le \tilde g_t(a)$ and $h_t(a) \le \tilde h_t(a)$, because $\Phi \le \tilde\Phi$. Because

$$g_1(a) = \inf\{b\colon as + b \ge \Phi(s), \forall s \in I\},$$

and similarly for $\tilde g_1$, the definition of $\tilde\Phi$ shows that $\tilde g_1(a) \le g_1(a)$ and hence $\tilde g_1(a) = g_1(a)$.

By the definition of argmax as the largest point of maximum it follows that $\operatorname{argmax}_s\{\tilde\Phi(s) - as\} < t$ if and only if the maximum of $s \mapsto \tilde\Phi(s) - as$ on $[t, 1]$ is strictly less than the maximum before t, i.e. if and only if $\tilde h_t(a) < \tilde g_1(a)$. This implies that $h_t(a) < \tilde g_1(a) = g_1(a)$, which in turn is equivalent to $\operatorname{argmax}_s\{\Phi(s) - as\} < t$.

Thus $\operatorname{argmax}_s\{\tilde\Phi(s) - as\} < t$ implies that $\operatorname{argmax}_s\{\Phi(s) - as\} < t$. For the proof of the converse assume that $\operatorname{argmax}_s\{\Phi(s) - as\} < t$. Then there exists $t' < t$ such that $g_{t'}(a) \ge g_1(a)$, which implies that $\tilde g_{t'}(a) \ge g_1(a) = \tilde g_1(a)$. Thus the map $s \mapsto \tilde\Phi(s) - as$ possesses a point of maximum strictly left of t. It remains to be proved that the largest point of maximum is also strictly left of t.

Suppose that this is not the case. Then there are points t_1, t_2 with $t_1 < t \le t_2$ at which $a \mapsto \tilde\Phi(s) - as$ is maximal. This implies that $s \mapsto \tilde\Phi(s)$ is below the line $s \mapsto \tilde\Phi(t_1) + a(s - t_1) = \tilde\Phi(t_2) + a(s - t_2)$ on $[t_1, t_2]$. By concavity this can happen only if $\tilde\Phi$ is equal to this line, i.e. $\tilde\Phi$ is linear on $[t_1, t_2]$ with slope a. Every point in $[t_1, t_2]$ is then a point of maximum of $s \mapsto \tilde\Phi(s) - as$ and we may choose a larger value of $t_1 < t$, if necessary, to ensure that the map $s \mapsto \Phi(s) - as$ assumes its maximal value on I on the subinterval $[0, t_1]$, i.e. $\Phi(s) - as < \Phi(t_1) - at_1$ for all $s > t_1$ and $\Phi(s) - as \le \Phi(t_1) - at_1$ for all $s \in I$. Fix a point t_3 with $t_1 < t_3 < t_2$. By the upper semi-continuity of Φ ,

$$c\colon = \Phi(t_1) - at_1 - \max_{s\ge t_3}\bigl(\Phi(s) - as\bigr) > 0.$$

Consider now the line L through the points $(t_3, \tilde\Phi(t_1) + a(t_3 - t_1))$ and $(1, \tilde\Phi(t_1) + a(1 - t_1) - c)$, i.e. $L(s) = \tilde\Phi(t_1) + a(t_3 - t_1) + \alpha(s - t_3)$, for slope

$$\alpha = a - \frac{c}{1 - t_3} < a.$$

It is straightforward to verify that $L \ge \Phi$ on I. (Use $\alpha < a$ on $[0, t_3]$ and the definition of c on $[t_3, 1]$.) By the definition of $\tilde\Phi$ it follows that $L \ge \tilde\Phi$ as well. However, $L(t_2) < \tilde\Phi(t_2)$, a contradiction. ∎

If Φ is not upper semi-continuous, then we might first replace it by its smallest upper semi-continuous majorant $\bar\Phi$, which is defined as the infimum over all functions $f\colon I \to \mathbb{R}$ with $f \ge \Phi$ that are upper semi-continuous. Next we can compute the concave majorant $\tilde{\bar\Phi}$ of $\bar\Phi$. This is bigger than $\tilde\Phi$, of course, but only at the boundary points, in view of the following lemma.

LEMMA 4.2. *Let $\Phi: I \subset \mathbb{R}$ be arbitrary on the compact interval $I \subset \mathbb{R}$. Then $\bar{\tilde{\Phi}} = \tilde{\Phi}$ on the interior of I.*

Proof. As a concave function on I, the function $\tilde{\Phi}$ is automatically continuous on the interior of I. Therefore, its upper semi-continuous majorant is obtained by increasing its values at the boundary points, if necessary:

$$\bar{\tilde{\Phi}}(x) = \begin{cases} \tilde{\Phi}(x), & \text{if } x \in \mathring{I}, \\ \limsup_{y \to x, y \in \mathring{I}} \tilde{\Phi}(y), & \text{if } x \in \delta I. \end{cases}$$

(Cf. Lemma 4.3 below. Note that for $x \in \delta I$ the value $\tilde{\Phi}(x)$ cannot be strictly bigger than $\lim_{y \to x, y \in \mathring{I}} \tilde{\Phi}(y)$, because of concavity. This is the reason that we can restrict the limit over y to y in the interior of I.) We claim that this function is concave on I. Indeed, it is concave on the interior of I. If x is in the boundary of I and y is in the interior of I, then, for some $x_n \to x$ with $x_n \in \mathring{I}$ and every $0 < t < 1$,

$$\begin{aligned} (1-t)\bar{\tilde{\Phi}}(y) + t\bar{\tilde{\Phi}}(x) &= \lim_{n \to \infty} (1-t)\tilde{\Phi}(y) + t\tilde{\Phi}(x_n) \\ &\leq \lim_{n \to \infty} \tilde{\Phi}\big(ty + (1-t)x_n\big) \\ &= \tilde{\Phi}\big(ty + (1-t)x\big) = \bar{\tilde{\Phi}}\big(ty + (1-t)x\big). \end{aligned}$$

Similarly, if both x and y are in the boundary of I. Hence $\bar{\tilde{\Phi}}$ is a concave, upper semi-continuous function that is bigger than Φ on I. Because $\tilde{\Phi} \geq \Phi$, it follows that $\bar{\tilde{\Phi}} \geq \bar{\Phi}$ Being concave, $\bar{\tilde{\Phi}}$ is then also bigger than $\tilde{\bar{\Phi}}$, so that $\bar{\tilde{\Phi}} \geq \tilde{\bar{\Phi}} \geq \tilde{\Phi}$. As we have seen, the leftmost and rightmost function are equal on the interior of I and hence the three functions coincide there. ∎

If Φ is a stochastic process, then so is its smallest upper semi-continuous majorant $\bar{\Phi}$. It is a fortunate fact that some important properties of a sequence of processes Φ_n are inherited by the sequence $\bar{\Phi}_n$. To derive this we need the following characterization of $\bar{\Psi}$.

LEMMA 4.3. *Let $\Phi: T \to \mathbb{R}$ be an arbitrary function on a metric space T. Then $\bar{\Phi}(t) = \limsup_{s \to t} \Phi(s)$ for every t.*

Proof. It suffices to show that the function Ψ defined by taking the limsup over Φ is upper semi-continuous, because we certainly have that $\bar{\Phi}(t) \geq \limsup_{s \to t} \bar{\Phi}(s)$ (because $\bar{\Phi}$ is upper semi-continuous by definition), which is bigger than $\Psi(t)$ (because $\bar{\Phi} \geq \Phi$). Now

$$\limsup_{s \to t} \Psi(s) = \inf_{\varepsilon > 0} \sup_{d(s,t) < \varepsilon} \inf_{\delta > 0} \sup_{d(u,s) < \delta} \Phi(u) \leq \sup_{d(u,t) < \eta} \Phi(u),$$

for any $\eta > 0$, as we can see by choosing $\varepsilon = \delta = \eta/2$. Then the left side is also smaller than the infimum of the right side over $\eta > 0$, which is $\Psi(t)$. ∎

LEMMA 4.4. *Let Z_n and Z be maps from a probability space into $\ell^\infty(T)$ for some metric space T such that Z is Borel measurable and separable and such that $Z_n \rightsquigarrow Z$ in $\ell^\infty(T)$. Then $\bar{Z}_n \rightsquigarrow \bar{Z}$ in $\ell^\infty(T)$ (provided that $\bar{Z}$ is Borel measurable).*

Proof. By the almost sure representation theorem (van der Vaart and Wellner (1996, Theorem 2.10.4), there exists a probability space and maps $\check{Z}_n, \check{Z}: \check{\Omega} \to \ell^\infty(T)$ such that $\check{Z}_n \to \check{Z}$ (outer) almost surely in $\ell^\infty(T)$ and, for every $n \in \mathbb{N}$ and every bounded function $f: \ell^\infty(T) \to \mathbb{R}$,

$$\mathrm{E}^* f(\check{Z}_n) = \mathrm{E}^* f(Z_n), \qquad \mathrm{E}^* f(\check{Z}) = \mathrm{E}^* f(Z).$$

For any given random variables ε_n the inequality $\|\check{Z}_n - \check{Z}\|_T \le \varepsilon_n$ is equivalent to $\check{Z}(t) - \varepsilon_n \le \check{Z}_n(t) \le \check{Z}(t) + \varepsilon_n$ for every $t \in T$, by Lemma 4.3, and implies that $\bar{\check{Z}}(t) - \varepsilon_n \le \bar{\check{Z}}_n(t) \le \bar{\check{Z}}(t) + \varepsilon_n$ for every $t \in T$. We conclude that $\bar{\check{Z}}_n \to \bar{\check{Z}}$ outer almost surely in $\ell^\infty(T)$. Because the function $g: z \mapsto \bar{z} \mapsto f(\bar{z})$ from $\ell^\infty(T)$ to $\mathbb{R}$ is bounded for every bounded function $f: \ell^\infty(T) \to \mathbb{R}$, we obtain that $\mathrm{E}^* f(\bar{Z}_n) = \mathrm{E}^* g(Z_n) = \mathrm{E}^* g(\check{Z}_n) = \mathrm{E}^* f(\bar{\check{Z}}_n) \to \mathrm{E}^* f(\bar{\check{Z}}) = \mathrm{E}^* g(\tilde{Z}) = \mathrm{E}^* g(Z) = \mathrm{E}^* f(\bar{\,}) = \mathrm{E}^* f(\bar{Z})$ for every bounded, continuous f, whence the result. ∎

Weak convergence in $\ell^\infty(T)$ is connected to the modulus of continuity. We can also compare the moduli of a stochastic process and its upper semi-continuous majorant directly.

LEMMA 4.5. *For any function $\Phi: T \to \mathbb{R}$ on a metric space T, $\delta > 0$, and open set $G \subset T$,*

$$\sup_{d(s,t)<\delta} \bigl|\bar{\Phi}(s) - \bar{\Phi}(t)\bigr| = \sup_{d(s,t)<\delta} \bigl|\Phi(s) - \Phi(t)\bigr|,$$
$$\sup_{t \in G} \bar{\Phi}(t) = \sup_{t \in G} \Phi(t).$$

Proof. This is an immediate consequence of Lemma 4.3. ∎

5. Covering Numbers

In this section we collect some preservation properties of uniform covering numbers, useful to bound the entropy of classes of complicated functions by bounds on the entropies of more standard classes.

Covering numbers $N\bigl(\varepsilon, \mathcal{F}, L_r(Q)\bigr)$ yield a measure of the complexity of a class $\mathcal{F}$ of measurable functions $f\colon \mathcal{X} \to \mathbb{R}$ relative to the L_r-norm corresponding to a measure Q. They are defined as the minimal number of balls of radius $\varepsilon > 0$ needed to cover $\mathcal{F}$. An envelope function F is a measurable function $F\colon \mathcal{X}| \to \mathbb{R}$ such that $|f| \le F$ for every $f \in \mathcal{F}$. We denote by $\|f\|_{Q,r}$ the norm of f in $L_r(Q)$.

The *uniform (L_2-)covering number* of a class $\mathcal{F}$ relative to the envelope function F is the number

$$\sup_Q N\bigl(\varepsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q)\bigr),$$

where the supremum is taken over all discrete probability measures Q on $(\mathcal{X}, \mathcal{A})$ such that $\|F\|_{Q,2} > 0$. The logarithm of this number is the uniform entropy of the class. A basic inequality using these numbers, due to Pollard (1989), is that

$$\mathrm{E} \sup_{f\in\mathcal{F}} |\mathbb{G}_n f| \lesssim \int_0^\infty \sqrt{\log \sup_Q N\bigl(\varepsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q)\bigr)}\, d\varepsilon\, (PF^2)^{1/2}, \tag{5.1}$$

provided the class $\mathcal{F}$ meets certain measurability requirements. (See Van der Vaart and Wellner (1996, Lemma 2.14.1).)

Given classes $\mathcal{F}_\infty, \ldots, \mathcal{F}_k$ of functions $f_i\colon \mathcal{X} \to \mathbb{R}$ and a map $\phi\colon \mathbb{R}^k \to \mathbb{R}$, let $\mathcal{F} = \mathcal{F}_\infty \times \cdots \times \mathcal{F}_k$ and let $\phi \circ \mathcal{F}$ denote the set of functions $x \mapsto \phi\bigl(f_1(x), \ldots, f_k(x)\bigr)$ as $f = (f_i, \ldots, f_k)$ ranges over $\mathcal{F}$. Assume that for some measurable functions $L_i\colon \mathcal{X} \to \mathbb{R}$, constants $\alpha_i \in (0, 1]$, every $x \in \mathcal{X}$, and every $f, g \in \mathcal{F}$,

$$\bigl|\phi(f(x)) - \phi(g(x))\bigr| \le \sum_{i=1}^k L_i(x)\bigl|f_i(x) - g_i(x)\bigr|^{\alpha_i}.$$

Then the function $2L \cdot F^\alpha := 2\sum_{i=1}^k L_i F_i^{\alpha_i}$ is an envelope function of the class $\phi(\mathcal{F}) - \phi(f_0)$, for any fixed $f_0 \in \mathcal{F}$, if F_i is an envelope function for $\mathcal{F}_i$. The function

$$(L \cdot F^\alpha)_r := \bigl(\sum L_i^r F_i^{r\alpha_i}\bigr)^{1/r}$$

is almost as good and more natural when using the L_r-norm ($r \ge 1$). We note that $k^{-1}L \cdot F^\alpha \le k^{-1/r}(L \cdot F^\alpha)_r \le k^{-1/r} L \cdot F^\alpha$, so that for fixed k these envelopes are equivalent.

A simple case of interest is when ϕ is uniformly Lipschitz of Lipschitz constant 1. Then we can take $\alpha_i = 1$ and $L \cdot F = \sum_{i=1}^k F_i$.

LEMMA 5.1. *For any $r \geq 1$,*

$$\sup_Q N\big(\varepsilon\|(L\cdot F^\alpha)_r\|_{Q,r}, \phi(\mathcal{F}), L_r(Q)\big) \leq \prod_{i=1}^k \sup_Q N\big(\varepsilon^{1/\alpha_i}\|F_i\|_{Q,r\alpha_i}, \mathcal{F}_i, L_r(Q)\big).$$

Proof. This is an extension and restatement of Theorem 2.10.20 of van der Vaart and Wellner (1996, page 199). Also see Pollard (1991, Section 5). ∎

Example (Products). For any $r \geq 1$ and any classes of functions $\mathcal{F}$ and G with envelopes F and G we have

$$\sup_Q N\big(\sqrt{2}\varepsilon\|FG\|_{Q,r}, \mathcal{F}\mathcal{G}, L_r(Q)\big) \leq \sup_Q N\big(2^{1/r}\varepsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)\big) \sup_Q N\big(\varepsilon\|G\|_{Q,r}, \mathcal{G}, L_r(Q)\big).$$

This follows from the lemma with $\phi(f,g) = fg$, $\alpha_1 = \alpha_2 = 1$, $L_1 = G$ and $L_2 = F$. We use this example in particular with $\mathcal{G}$ consisting of the single function G, in which case of course the last supremum in the display is equal to 1.

Example (Sums). For any $r \geq 1$ and any classes of functions $\mathcal{F}$ and G with envelopes F and G we have

$$\sup_Q N\big(\varepsilon\|F+G\|_{Q,r}, \mathcal{F}+\mathcal{G}, L_r(Q)\big) \leq \sup_Q N\big(\varepsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)\big) \sup_Q N\big(\varepsilon\|G\|_{Q,r}, \mathcal{G}, L_r(Q)\big).$$

This follows from the lemma with $\phi(f,g) = f+g$, $\alpha_1 = \alpha_2 = 1$, and $L_1 = L_2 = 1$.

Example (Quotients). For any class $\mathcal{F}$ of functions $f\colon \mathcal{X} \to [\eta, \infty)$ for some $\eta > 0$, we have

$$\sup_Q N\Big(\varepsilon\Big\|\frac{F}{\eta^2}\Big\|_{Q,r}, 1/\mathcal{F}, L_r(Q)\Big) \leq \sup_Q N(\varepsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q)\big).$$

This follows from the lemma with $\phi(f) = 1/f$ and $L_1 = 1/\eta^2$. The boundedness from below can be much relaxed, by using a lower envelope function, along the lines of Example 2.10.22 in Van der Vaart and Wellner (1996).

Next consider functions formed by integrating out one variable from a set of functions of two variables. Given a measurable function $f\colon \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ on a product measurable space and a fixed probability measure R, let

$$\bar{f}(x) = \int f(x,y)\, dR(y).$$

Let $\bar{\mathcal{F}}$ be the set of all functions $\bar{f}$ as f ranges over $\mathcal{F}$. The function $\bar{F}$ for F an envelope of $\mathcal{F}$ is a natural envelope for this class. When using the L_r-norm the function

$$\bar{F}_r(x) = \big(\int F^r(x,y)\, dR(y)\big)^{1/r},$$

which is bigger in general for $r \geq 1$ by Hölder's inequality, is sometimes more natural.

LEMMA 5.2. *For any $r \geq s \geq 1$ and $t \geq s$,*

$$\sup_Q N\big(2\varepsilon\|\bar{F}_s\|_{Q,r}, \bar{\mathcal{F}}, L_r(Q)\big) \leq \sup_Q N\big(\varepsilon^{r/s}\|F\|_{Q,t}, \mathcal{F}, L_t(Q)\big).$$

Proof. This is an interpolation between results given by Sherman (1994, Lemma 5) and Ghosal, Sen and Van der Vaart (2000, Lemma A2). Because the use of the parameter s is essential for this paper and makes the result somewhat non-trivial, we include the proof.

By Jensen's inequality, for any probability measure Q,

$$Q|\bar{f} - \bar{g}|^s \leq (Q \times R)|f - g|^s.$$

Furthermore, $Q\bar{F}_s^s = (Q \times R)F^s$. Hence, for every $\varepsilon > 0$ and Q,

$$N(\varepsilon\|\bar{F}_s\|_{Q,s}, \bar{\mathcal{F}}, L_s(Q)) \leq N(\varepsilon\|F\|_{Q\times R,s}, \mathcal{F}, L_s(Q \times R)).$$

The right side does not decrease if the L_s-norm is replaced by the L_t-norm for $t \geq s$, by Problem 2.10.4 of van der Vaart and Wellner (1996). (We note that the condition that the envelope function be strictly positive in that problem is superfluous: the covering number $N\big(\varepsilon\|F\|_{Q,s}, \mathcal{F}, L_s(Q)\big)$ for an arbitrary measure Q with $QF^s > 0$ is not bigger than $N\big(\varepsilon\|F\|_{Q_1,s}, \mathcal{F}, L_s(Q_1)\big)$ for Q_1 the probability measure defined by $Q_1(A) = Q(A \cap \{F > 0\})/Q(\{F > 0\})$.) After this replacement the right side of the preceding display is bounded by the right sice of the lemma with ε instead of $\varepsilon^{r/s}$.

We now conclude the proof by showing that

$$\sup_Q N\big(2^{1-s/r}\varepsilon\|\bar{F}_s\|_{Q,r}, \bar{\mathcal{F}}, L_r(Q)\big) \leq \sup_Q N\big(\varepsilon^{r/s}\|\bar{F}_s\|_{Q,s}, \bar{\mathcal{F}}, L_s(Q)\big).$$

To see this note first that

$$Q|\bar{f} - \bar{g}|^r \leq Q|\bar{f} - \bar{g}|^s(2\bar{F}_s)^{r-s} = 2^{r-s}P|\bar{f} - \bar{g}|^s Q\bar{F}_s^{r-s},$$

for the measure P defined by $Pf = Qf\bar{F}_s^{r-s}/Q\bar{F}_s^{r-s}$. Therefore, if $P|\bar{f} - \bar{g}|^s \leq \varepsilon^r\|\bar{F}_s\|_{P,s}^s$, then $Q|\bar{f} - \bar{g}|^r \leq 2^{r-s}\varepsilon^r Q\bar{F}_s^r$. This proves the desired inequality. ∎

The appearance of the power $\varepsilon^{r/s}$ in the right side of the first inequality of the preceding lemma is disconcerting. It is ε^r when using the smaller, "natural" envelope function $\bar{F} = \bar{F}_\infty$, but reduces to ε when using the bigger envelope function $\bar{F}_r$. If applied to the class $\mathcal{F}k_b$ of functions $(x, y) \mapsto f(x, y)k(y/b)/b$ the inequality with $r = s = t = 2$ gives the envelope function $\bar{F}_2(x) = \big(\int F^2(x, y)k^2(y/b)/b^2\, dR(y)\big)^{1/2}$, which is of the order $1/\sqrt{b}$ if $F = 1$ and R is the uniform measure on $[-1, 1]$. Thus this type of envelope function behaves badly if the bandwidth b tends to zero. On the other hand the natural envelope $\bar{F}(x) = \int F(x, y)k(y/b)/b\, dR(y) = 1$ (if $F = 1$) is bounded, as it should be for this class of functions. The parameter s allows to make a trade-off between a "bad envelope" and a "bad power of ε". If the class $\mathcal{F}$ is "small", then its entropy evaluated at ε^2 rather than ε will still be small, and we can use the good envelope function. For a large class using the bad envelope may be preferable, in particular if the entropy of $\mathcal{F}$ is large than $1/\varepsilon$.

6. Acknowledgements

We thank Geurt Jongbloed for making available some of the results on concave majorants given in Section 4.

REFERENCES

[1] Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N., (1992). *Statistical Models Based on Counting Processes*. Springer, Berlin.

[2] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A., (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

[3] Gill, R.D., van der Laan, M.J. and Robins, J.M., (1997). Coarsening at Random: Characterizations, Conjectures and Counter-examples. Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis, 255–295, (eds: D.Y. Lin and T.R. Fleming). Springer Verlag.

[4] Groeneboom, P., (1985). Estimating a monotone density. Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer **2**, 539–555, (eds: L.M. Le Cam and R.A. Olshen). Wadsworth, Monterey, California.

[5] Groeneboom, P., (1987). Asymptotics for interval censored observations. Report **87-18**. Department of Mathematics, University of Amsterdam.

[6] Groeneboom, P. and Wellner, J.A., (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser, Basel.

[7] Heitjan, D.F. and Rubin, D.B., (1991). Ignorability and coarse data. *Annals of Statistics* **19**, 2244–2253.

[8] Huang, J., (1996). Efficient estimation for the Cox model with interval censoring. *Annals of Statistics* **24**, 540–568.

[9] Jacobsen, M. and Keiding, N., (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *Annals of Statistics* **23**, 774–786.

[10] Jongbloed, G., (1999). *Inverse Problems in Statistics*. Lecture Notes, Vrije Universiteit.

[11] Kim, J. and Pollard, D., (1990). Cube root asymptotics. *Annals of Statistics* **18**, 191–219.

[12] van der Laan, M.J. and Hubbard, A., (1997). Estimation with Interval Censored Data and Covariates. *Lifetime Data Analysis* **3**, 77–91.

[13] Murphy, S.A. and van der Vaart, A.W., (1997). Semiparametric likelihood ratio inference. *Annals of Statistics* **25**, 1471–1509.

[14] Murphy, S.A. and van der Vaart, A.W., (1999). Observed information in semiparametric models. *Bernoulli* **5**, 381–412.

[15] Murphy, S.A. and van der Vaart, A.W., (2000). On Profile Likelihood. *Journal of the American Statistical Association* **95**, 449–465.

[16] Pollard, D., (1989). Asymptotics via empirical processes. *Statistical Science* **4**, 341–366.

[17] Pollard, D., (1990). *Empirical processes: theory and applications*. *NSF-CBMS Regional Conference Series in Probability and Statistics* **2**. Institute of Mathematical Statistics and American Statistical Association, Hayward.

[18] Robertson, T., Wright, F.T. and Dykstra, R.L., (1988). *Order Restricted Statistical Inference*. Wiley, New York.

[19] Robins, J.M., (1993). Information recovery and bias adjustment in proportinal hazards analysis of randomized trials using surrogate markers. Proceedings of the Biopharmaceutical Section, 24–33. American Statistical Association.

[20] Robins, J.M. and Rotnitzky, A., (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS Epidemiology–Methodological Issues*, 297–331, (eds: N. Jewell, K. Dietz, and V. Farewell). Birkhäuser, Boston.

[21] Robins, J.M. and Ritov, Y., (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* **16**, 285–319.

[22] Robins, J.M. and van der Laan, M.J., (1998). Locally efficient estimation with current status data and time-dependent covariates. *Journal of the American Statistical Association* **93**, 693–701.

[23] Sherman, R.P., (1994). Maximal inequalities for degenerate U-statistics with applications to optimization estimators.. *Annals of Statistics* **22**, 439–459.

[24] van der Vaart, A.W., (1991). On differentiable functionals. *Annals of Statistics* **19**, 178–204.

[25] van der Vaart, A.W., (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

[26] van der Vaart, A.W., (2001). On a theorem by James Robins. *preprint,*.

[27] van der Vaart, A.W. and Wellner, J.A., (1996). *Weak Convergence and Empirical Processes*. Springer, New York.